DOCUMENT RESUME

ED 072 228	VT 018 606

AUTHOR	Creager, John A.
TITLE	Noneconomic Analysis Considerations for Management and Information System for Occupational Education.
INSTITUTION	Management and Information System for Occupational Education, Winchester, Mass.
PUB DATE	15 Jun 72
NOTE	136p.; Occasional Paper-7
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Algorithms; Design Preferences; Economic Research; Educational Planning; Information Retrieval; *Management Information Systems; Man Machine Systems; *Mathematical Models; *Operations Research; Simulation; *Statistical Analysis; Systems Development; *Vocational Education
IDENTIFIERS	*Management Information System Occupational Educa; Massachusetts; MISOE

ABSTRACT

As the first of two papers delineating the design of Massachusetts' Management and Information System for Occupational Education (MISOE), these specific dimensions of MISOE structure and function are considered: (1) the distinction between economic and noneconomic analysis, (2) distinctions among census, sample, and other data, (3) the distinction between descriptive and simulative analysis, and (4) functional levels, management levels, and management scope. Information retrieval and analysis for MISOE necessitates: (1) translation of inquiries into analytic hypotheses, (2) the selection of pertinent MISOE subsystems, data types and levels, analytical operations, and models, (3) performing the analyses and interpreting their results, and (4) reporting the results to the inquiry source. Discussions of general analysis requirements and considerations precede the detailing of specific analytical models and algorithms for MISOE, such as multiple linear regression and factor analysis. Dynamic simulation, linear programing, and nonlinear programing models are discussed, in addition to specific noneconomic analysis factors to consider within and among MISOE's subsystems of static space. Technical reports on MISOE's research methodology are appended. Related documents are available in this issue as VT 018 600, VT 018 602, VT 018 809, and VT 018 810. (AG)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 072228

OCCASIONAL PAPER #7

NONECONOMIC ANALYSIS CONSIDERATIONS FOR MISOE

By

John A. Creager

June 15, 1972

VT 018606

Management and Information System for Occupational Education
1017 Main Street, Winchester, Massachusetts 01890
Telephone: 617-729-9260

Preface

This paper is one of a series prepared by the staff and a team of consultants to delineate and document the design of the Management and Information System for Occupational Education. It is the first of two such papers by the author, submitted as the formal response to staff inquiries, and as major tangible products of the consultation relationship. Gratitude is expressed to the staff for its extensive help, in documentation and in conferences. Although all reasonable effort has been made to be both relevant and accurate, the author disclaims infallability, and encourages the staff to be both selective and flexible in its use of the aids offered.

TABLE OF CONTENTS

PART I. General Definition of Context, Scope, and Depth of Analysis Considerations........ 1

A. The Context........ 1

B. General Analysis Requirements........ 2

C. "Dimensions" for Analysis Requirements........ 3

PART II. General Analysis Considerations........ 5

Sources and Control of Inferential Errors........ 5

General Software Considerations........ 9

Missing Data Problems........ 12

Formulation of "Mixes"........ 13

Followup Problems........ 14

Implications of "Initial Data Points" and Cohort Replacement........ 16

Some Functional Issues in Data Processing and the Computer Facility........ 17

PART III. The Analysis Repertoire of MISOE: Models and Algorithms........ 20

Multiple Linear Regression........ 20

Relative Contributions in Regression........ 23

The Control of Process-Product and Process-Impact Analyses for Differential Input........ 28

Principal Components and Factor Analysis........ 29

Canonical and Discriminant Analysis........ 30

Temporal Analysis........ 32

Miscellaneous Analysis Tools of Moderate to Lower Priority........ 33

PART IV. Noneconomic Analysis Considerations Within and Among Subsystems of Static Space........ 36

Introduction........ 36

The Range Restriction Problem........ 36

Analysis Considerations for Noneconomic Factors in the Process Space........ 40

Analysis Involving the Product Space........ 45

Analysis Involving the Impact Space........ 51

The Educational Human Input and Student Spaces........ 55

Analysis Across IPPI Spaces........ 57

TABLE OF CONTENTS
(Continued)

PART V. Simulation Models 59

Introduction 59

General Consideration of Dynamic Simulation 62

Equations and Data Sources.................................. 64

Inferential Errors in Dynamic Simulation 69

A Pseudodynamic Model as Nonlinear Programming.............. 73

Figure I. Pseudodynamic Model for Process-Product Inquiry..... 75

A More Rational Approach.................................... 77

A Linear Programming Solution 80

PART VI. Epilogue.. 83

APPENDIXES Appendix A Technical Reports for MISOE i

Appendix B Memoranda for MISOE xl

NONECONOMIC ANALYSIS CONSIDERATIONS FOR MISOE

John A. Creager

Part I. General Definition of Context, Scope, and Depth of Analysis Considerations

A. The Context

The general purposes, structures, and functions of MISOE have been delineated in Monograph No. 1, Occasional Papers 1-6, and in a position paper. Although MISOE has primary reference to occupational education in the State of Massachusetts, it is recognized that occupational education is imbedded in the general state system of education, which in turn is imbedded in the still more general system of state concerns for realizing societal values. Moreover, MISOE is to be prototypical, i.e., a paradigm for other management and information systems, and therefore, a contribution to management technology as well as a practical management tool for occupational education in Massachusetts.

In the anticipated typical usage of MISOE, an inquiry will be initiated at some management level and translated into a problem for information retrieval and analysis. The resulting information and its implications by interpretation will then be fed back to the source of inquiry as a system response. Where the initial inquiry is complex, it may be fractionated into subinquiries, each of which demand information retrieval and analysis, with their implications integrated into the MISOE response to the inquiry source. An inquiry (or subinquiry) implies:

1. translation into analytic hypotheses
2. selection of the relevant MISOE subsystems (spaces and elements)
3. selection of the relevant data types and levels
4. selection of the relevant models and analytical operations
5. performance of the analyses
6. interpretation of analysis results
7. reporting of interpreted results to inquiry source.

An inquiry may involve economic information (e.g., costs, time, etc.), noneconomic information (quality and quantity of manpower, educational programs, etc.), or both. Occasional Papers No. 7 and No. 9 are addressed to considerations of the noneconomic analysis aspects of MISOE, while Occasional Paper No. 8 is addressed to the consideration of economic analysis. Because complex inquiries, especially those received from the state management level, are likely to involve both economic and noneconomic aspects in relation to each other, more intensive attention must be given in further MISOE development to integrating these aspects in analysis. This paper assumes that such interrelations will occur analytically in many of the simulation analyses rather than in the descriptive (nonsimulation) analyses, but it is recognized that some of the economic-noneconomic relationships will be descriptive information useful in formulating auxiliary equations that moderate flow rates in dynamic space. Although this paper is focused on noneconomic analysis considerations, some further consideration of these matters will be made in a later section (in Part FIVE) on the interface and communication between descriptive and simulation analysis.

B. General Analysis Requirements

The general requirements for the analytic aspects of MISOE must recognize the demands that MISOE have:

1. generality in terms of level and scope of inquiry,
2. flexibility in terms of changes in educational programs, available technology, variations in inquiry types,
3. expandability in terms of new programs, new issues and inquiries, and new data types, and enlarged capability to simulate finer aspects of "reality",
4. continuity of operational capability regardless of personnel or other changes, and
5. sensitivity to the needs of potential inquirers of the system.

Moreover, the analytic aspects of MISOE must interface and intercommunicate all aspects of MISOE structures and functions. Thus, the analytic capabilities contribute:

1. to generality by MISOE having in its repertoire general models and computer programs,
2. to flexibility by having a broad repertoire of such models and programs with many specific options,
3. to expandability by being aware of analytical tools not immediately required, but of potential value as MISOE expands,
4. to continuity by having thorough documentation and referencing of all models, computer programs, and analyses actually performed, and
5. to sensitivity to potential user needs by the scope and depth of the MISOE analysis repertoire.

C. "Dimensions" for Analysis Requirements

This section defines some major "dimensions" of MISOE structure and function. Analytic considerations which are general and therefore cut across such dimensions are discussed in Parts Two and Three. Part Two discusses those topics which cut across the models and algorithms, which constitute the analysis repertoire and which are discussed in Part Three. The present discussion of "dimensions" defines the categories on each dimension and how they are treated.

The first dimension is the distinction between economic and noneconomic analysis requirements and was discussed above. In terms of the analysis types stated in Occasional Paper No. 1, the emphasis in this paper on noneconomic analysis implies emphasis on A2 (process-product), A4 (product-impact), and A5 (process-impact).

The second dimension distinguishes census, sample, and other data. This paper assumes that all analysis above the entry level (and some of it there) is concerned with the census populations, or subcategories thereof, and that sample data will be appropriately weighted to be representative of those populations. The requirements to ensure that this is so for all data types and analysis levels will be delineated in Occasional Paper No. 12, which will specify sampling and weighting procedures. Data from external sources, e.g., U.S. Census, Project CAREER, Project TALENT, the Cooperative Institutional Research Program (CIRP) of the American Council on Education (ACE), will be only partially connectable for comparative and normative information at analysis levels 1 and 2 (see Occasional Paper No. 2).

The distinction between descriptive and simulative analysis constitutes a third dimension. Descriptive analysis includes the estimation of population parameters from sample data (discussed for entry level analysis in Occasional Paper No. 12), distributional statistics and correlational analysis both univariate and multivariate, and special capabilities for discrimination, data transformation, and taxonomy. Simulative analysis includes both static simulation, which may be required by certain types of inquiry, and dynamic simulation of the Forrester type. Analysis considerations will be discussed both within and between the descriptive and simulative categories.

The discussion of descriptive analytic considerations will also be structured in terms of the spaces and elements of MISOE delineated in Occasional Papers No. 1, No. 2, and No. 4, as expanded and modified in Occasional Paper No. 5. Part Four discusses analysis within and across such MISOE subsystems. These discussions will take cognizance of the dimension of functional levels (educational sectors such as secondary, adult, MDT, programs, blocks, and units), the dimension of management level (state, region, or local), and the dimension of management scope (social agencies, all education, occupational education across programs, and occupational education within programs). Generally, however, it is assumed that beyond the entry level, where the information storage and retrieval system (Occasional Paper No. 3) identifies such information, and where aggregation will be accomplished, the nature of a particular inquiry will imply the functional level and analysis units without need for major operating decisions in connection with choices of models and other analysis tools.

Part II: General Analysis Considerations

This part considers analysis topics which are not specific to particular inquiries or their associated relevant MISOE subsystems. Thus, they are likely to be involved in any analytic operations to some degree. Included are concerns about analysis tools and other such topics as sources and control of inferential errors, computer software, and special problems involving "mixes", cohort replacement, and followups. General analysis considerations also include the selection of variables and data instruments, a topic which will not be discussed here because it is the subject of Occasional Paper No. 10.

Sources and Control of Inferential Errors

The general utility of MISOE is to be that of a management tool for the appraisal of existing policies and as a guide to decision-making and policy change at all management levels. A system of this complexity contains a number of hazards that may arise anytime during operation from initial inquiry to the final interpreted response of the system. This section is concerned with such hazards that arise in analysis operations and which may distort the results of analysis in such a way that a false or misleading picture of "reality" is inferred (inferential error). The recommendations to management may then be faulty and even dangerous, either to individuals in the state system under study, to programs, to the managers who believe and act on the recommendations from MISOE, or ultimately, to MISOE itself. In addition to the obvious issue of professional ethics involved, the credibility and therefore the survival of MISOE is at stake.

The major sources of inferential error in analysis arise from sampling error, measurement error, and processing error. Errors may be either random or nonrandom in either their occurrence, or in their analytic consequences. Much of the classical literature in psychometrics and the other behavioral sciences has limited applicability in complex programs using data obtained from diversified sources and over extended periods of time.

The sources and control of sampling error will be discussed in somewhat greater depth in Occasional Paper No. 12. Suffice it to say here that:

1. random sampling errors are controlled by sample size and by stratification;
2. nonrandom sampling errors are controlled partly by the logistics of data collection, and partly by stratification;
3. the effects of both kinds of error are partially controllable by suitable procedures for weighting data to make them more nearly representative of the populations of interest;
4. bias from nonrandom sampling, including that from nonresponse to mail or phone contact of subjects, is a much more serious concern, and more difficult to control, than the effects of random sampling fluctuations.

Concern about the reliability of measurements, i.e., the consistency of data obtained on replicated measurement, pervades the classical literature in psychometrics. It is usually not difficult to obtain good estimates of reliability of instruments in common use and of good repute, especially for the standard tests of ability and achievement. For the most part such instruments have sufficiently high reliability to ensure the usefulness of the instrument, and in cases where coefficients are not very high, special procedures are available to reduce the likelihood of inferential error. In large scale programs (TALENT, CIRP of ACE, and MISOE), where many different kinds of measurements may enter a particular analysis, these measures having varying degrees of reliability, and where the reliabilities may also vary with respect to various groups of subjects, the risks of inferential errors from measurement error may interact with one another in such a manner as to render the results quite uninterpretable. In the case of categorical variables used to define subgroups of the population under study, measurement error leads to misclassifications. Relations between such variables and other variables are not necessarily attentuated but may be spuriously inflated. In regression analysis, different reliabilities across the variables may not only disturb the relative size of regression weights, but may even reverse their relative order of magnitude. This can be serious where such weights are later used as auxiliary modifiers of rates in dynamic simulation, or used in rendering some judgement about the relative importance of the regression variables in prediction, or used in accounting for variance in a dependent variable. The ACE Research Report, Measurement Error in Social and Educational Survey Research, deals with many of the issues raised here, and should be regarded for MISOE development purposes as an appendix to this Occasional Paper. Topics discussed include:

1. the meaning of measurement error,
2. the effects of error on analysis and interpretation,
3. a review of the pertinent literature (and a rather extensive bibliography),
4. concern about sources of error in different item types, formats, and contents,
5. error in those initial data-processing operations which have the same effects in analysis as measurement error,
6. empirical data on reliability of a variety of questionnaire survey item types.

Table 3 in Occasional Paper No. 5 shows staff concern for the reliability of instruments. Unless there is some sole available measure of a very important variable, it should not be necessary for MISOE to deal with any continuous variable with internal consistency or retest r_{ii} less than .75, even in the personality and attitude domains. Correlations involving continuous variables with r_{ii} in the .75-.95 range should be corrected for attenuation. This correction can be ignored when r_{ii} is greater than .95. Although somewhat arbitary, and set a little higher than necessary for exploratory research purposes, these cutpoints and associated recommendations are made to reduce the risk of letting measurement error have undue influence on outcomes from descriptive analysis upon parameters in dynamic simulation. In the case of dichotomous categorical variables, phi coefficients used as measures of reliability are a function of the base rate or popularity of the item and their values are constrained to a range of less than 0-1. When used to judge the reliability of the variable, phi should be divided by the maximum value that Phi can have for the associated base rate. However, point-biserial phi coefficients as measures of correlations between two variables should not be corrected for base rate, nor should point-biserial correlations between dichotomous and continuous variables, when these coefficients are used in correlational analysis (the use of normal biserial or tetrachoric coefficients is not recommended). Further discussion of reliability issues will be made when required in the later and more detailed discussions of analysis within and among MISOE subsystems and in simulation aspects of analysis (Parts Four and Five).

Errors occurring in the processing of information can also have serious inferential consequences. Processing includes the choice of models and algorithms to be discussed in Part Three. It is recommended that, as an integral part of MISOE development, some preliminary analyses using available or even fictitious data be run to thoroughly debug all computer software systems including such program modifications as may be required, to check out all information storage and retrieval aspects of the system including file manipulation and documentation, checking out the costs, time, and logistic aspects of operating MISOE, and the intercommunicability of all MISOE substructures. All of this is a part of "tooling up" and may save the staff much later embarrassment or even grief. Similar considerations apply to any later expansion of the system involving additional data processing operations.

The matter of documentation is not confined to data processing. In operation, an inquiry file should indicate not only the nature and source of the inquiry and the final response, but also any interim operations of analysis, so that accumulated experience can be referenced when faced with new inquiries and records of service experience can be used to train new personnel as the system expands or personnel changes occur.

General Software Considerations

This section is addressed to a few general considerations about the data processing and computer software requirements and capabilities of MISOE.

At the data entry level, test answer sheets, questionnaires, and other protocols must be processed in such a way that information can be read into various storage systems, which are either internal or external to the computer facilities. Where such documents require preliminary coding and/or keypunching, independent verification is strongly recommended. Staff consideration should also be given to the feasibility of using optical scanning and/or optical character reading operations. Such procedures are usually at least as reliable, and sometimes more so, than verified keypunching and result in the information from the document appearing on magnetic tape, as coded in accordance with user specifications. Moreover, the tape may contain data on variables that can be generated from the directly scanned responses, and additional "summary" tapes may be generated which contain aggregate data. Where feasible, the systems are quite flexible and provide much convenience for the user. Feasibility is a function of the volume and of the design of the protocol.

It is recommended that staff contact be made with representatives of the following to ascertain in more detail the feasibility of selective use of such services in MISOE operations at the data entry level:

Intran Corporation
4555 W. 77th Street
Minneapolis, Minnesota 55435
612-929-4691
Contact: Mr. Gerald Koch, President or Mr. Dennis Dillon

National Computer Systems
4401 W. 76th Street
Minneapolis, Minnesota 55435
612-920-6370
Contact: Dr. Robert J. Panos, Director of Survey Research Services

National Scanning Incorporated
1110 Morse Road
Columbus, Ohio 42339
Contact: Mr. Robert Hopkins, Director of Marketing.

The first two of these have been successfully used by ACE; the performance quality of NSI is less familiar. The first two companies use the same scanning principles based on transmitted light and require respondents to use a lead pencil (about no. 2½) for marking their responses. The representative of the last company claims that they can reliably read marks made with ball point or felt-tipped pens as well. Their system is based on a reflected light principle.

Software for the digital computer facility includes compilers and program packages. In addition to Fortran and Dynamo compilers, Cobol can be useful for file record counting and simple manipulations such as match-merge and pull-off of subfiles, and for certain kinds of accounting operations. Data-Text is a general program capable of generating more special programs for producing frequency distributions, distribution statistics and cross-tabulations in several dimensions; it is alleged to be imminently available in Fortran from

Harvard University Computing Laboratory (contact Dr. David Armour).

The system of analysis programs, known as BIOMED and available in Fortran from Dr. W. J. Dixon of UCLA, is very general and provides many options. One particular feature is the TRANSGENERATION option that permits variables to be transformed in terms of various functions, including crossmultiplication with other variables prior to entering a particular analysis algorithm. DYNAMO contains similar options.

Such general compilers and program packages, even when available, need to be adapted to the particular hardware configuration of the computing facility to be used by MISOE. The staff should ensure that these capabilities, including special routines for file sorting, handling multireel files, blocking and unblocking tapes, are well adapted, debugged, and documented, and the system monitor and all compilers contain thorough diagnostic capability.

With such software capability, the need for ad hoc programming should be minimal. Nevertheless, MISOE operations may encounter special requirements that imply special programming. Many of the subroutines in DYNAMO can be adapted, if necessary, to table lookup, plotting, and similar needs. BIOMED also provides some capabilities in these areas. It may also be anticipated that a user, upon studying a relationship plotted from empirical data and for which the function is not known, will need to fit a function to the plotted information, so that the equation or some of its parameters could be used in dynamic space. Such curvefitting capability may be rather important for MISOE. There will be a need for special programming for developing sampling weights and ensuring their additions to data records in such a way that the weights may be used selectively in analysis. This matter will receive further attention in a later section of this paper and in Occasional Paper No. 12.

Missing Data Problems

For a variety of reasons the data record for an observation unit may be incomplete. A student is absent on the date of testing; he omits some items on a questionnaire. In process and product spaces, local program records may not be complete or completely reported, although the staff may be able to elicit the missing information by phone. Incomplete records may be encountered in impact space (crime records not complete; followup respondents omit items, etc.). Where whole records are missing, sampling weights may need adjustment. This section, however, is concerned with sporadic losses of information (i.e., more or less random and not too frequent). On some sensitive items like family income, losses may run as high as 10 percent of the respondents. Where less than 1 percent losses are encountered, the analytic consequences may be negligible.

If the computer facility distinguishes zeros from blanks, the coding of data could use zero to code legitimate "zero", "nothing", "no", or "none" responses, and use "blank" for missing information. If the computer facility does *not* make such a distinction, or the staff wishes to ensure that files could be processed on external systems (e.g., for backup if the main facility is "down"), then zero should be used for missing data codes, dichotomous variables coded 2 or 1 (instead of 1 or 0), and all other coding of legitimate information subjected to a similar transformation to avoid using zero except to indicate missing data. Variances and covariances are unaffected by such transformation; only counts, aggregates (ΣX, ΣXY, and ΣX^2) and means are affected and are easily adjusted for reporting purposes. Where means are used to initiate level parameters in dynamic simulation, it is necessary to correct the level for the coding translation.

There are three general approaches to dealing with missing data in analysis. One is to leave the missing data coded as such on the files and to modify analysis programs to detect and bypass the code when cumulating sums. This is not recommended for MISOE. The other two methods replace missing information with estimated values. In one, an average value is computed across all records on the file for which data are available for the given variable. For categorical variables, the code for the modal category may be posted; for continuous variables means (sensitive to complete distribution assymetry) or medians (insensitive to assymetry) may be used. This procedure assumes that missing information is distributed symmetrically about the replacement value, an assumption that is usually false. The procedure also introduces some small attenuation in variance and covariances, but is a practical solution for MISOE, which can be accomplished during file editing operations. The final method is a refinement requiring considerable additional time and computing effort. Different replacement values are posted depending on the values of other, presumably related variables either by stratification on the other variables or by using regression estimates. This reduces the attenuation of variances and covariances and allows for assymetric losses of information from the total distribution. However, if missing data is sporadic and not too frequent, such refinements are probably unnecessary. Moreover, if the overall average is supplied for each file, when developed and edited, some stratification is implicitly introduced by separate files being developed from separate sources, for different MISOE subsystems, and levels.

The Formulation of "Mixes"

The use of the term "mix" in Occasional Paper No. 5 to denote certain patterns or configurations of student characteristics, process treatments, and product achievements raises issues with analytical implications. By far the most general and consistently useful formulation of a "mix" is the multivariate score vector ($X_1, X_2, \ldots, X_n$), which underlies most models for multivariate statistical analysis. In fact a data record (without its ID number and storage location information) is such a vector, any subset or linear transformation of which can express a "mix" analytically. Geometrically such a vector can be represented by a point in multivariate space, the mean vector is the vector of means (centroid of the swarm of points), and the space can be divided off into regions by various methods to provide codable groups of "similar mixes". (See Multivariate Statistics for Personnel Classification, Rulon, Tiedeman, Tatsouka, and Langmuir, 1967; the logic is also applicable to entities other than persons). All of which makes this the most attractive, general purpose formulation of "mixes" for analytic purposes, and applicable across the interfaces of MISOE subsystems. By this reasoning, the various clustering and "pattern analytic" methods (e.g., Tyron, Coombs) are not recommended for incorporation into the MISOE analytic capability, but in any case can be added later on, if need arises, as one kind of system expansion. The Guttman scaling approach discussed in Occasional Paper No. 5 may prove quite useful; this is discussed further in Part Four.

Any pattern can be coded and treated as an analytic datum regardless how that pattern was defined. The pattern variable is dichotomous: X takes one value if the pattern applies to the observation unit, another value if it does not. Such pattern coding in MISOE will probably be involved mostly at the higher analysis levels after factor analysis, discriminant analysis, or hierarchical grouping analyses have reduced the dimensions of the patterns; the number of possible patterns increases astronomically with the number of dimensions and the number of categories on each dimension.

Followup Problems

On completion of occupational education (or non-OE) a student moves back into the larger societal space and his behavior reflects the impact of his educational experience on him (product) and in turn, his impact on society as a taxpayer, producer, consumer, voter, or felon. The anticipation of assessing certain impact data through mail and/or phone follow contacts of "alumni requires that MISOE lay the groundwork for such contact while the student is still in the pipeline. The student's name and home address (or "address where he can always be reached") should be obtained on entry (so that dropouts can also be followed up) and updated on exit from educational programs. Provision should also be made during each followup inquiry for updating the address for the next followup of the same subject.

The name and address file should contain the subject's ID number, date of birth, and sex to help differentiate the James Jones's, those with such first names as Vivian or Shirley, and those with first names that might be confused through keypunch errors: Carl vs. Carol vs. Caroll, Marion vs. Marian, etc. The name and address file should, of course, be maintained separately from, and at a higher level of security than the coded data files. The use of separate ID numbers for data and name and address files with a link file provides additional confidentiality control at somewhat higher cost and staff operating inconvenience. An extensive literature has developed regarding confidentiality and related ethical and legal problems with data banks.

It would be prudent for the staff to maintain a backup copy of the name and address file and of each edited data file, whether a basic or followup file, and to keep the backup files in a separate location.

The creation of merged files containing basic and followup respondent data may be anticipated. This will involve special weighting procedures to adjust the data for bias due to nonresponse to the followups. These issues will be further delineated in Occasional Paper No. 12. The various analysis issues, including those involving interfacing MISOE subsystems, will apply to the followup data and operations on the merged files. Models and computing procedures will be similar to those for analysis of basic files except for changes of variables and their associations with MISOE subsystems.

Implications of "Initial Data Points" and Cohort Replacement

Rather than initiating MISOE solely with an input cohort in each process channel and following it through time (the purely longitudinal approach), which strategy requires waiting for product and impact data to become available, the staff has proposed combining longitudinal and cross-sectional designs. Thus, data will be collected initially not only in input space for an input cohort, but also in current process space and product space by whatever stage earlier cohorts may be in a given program, and in impact space for recent "alumni". The latter includes an initial prototype followup survey using addresses available in school records. It may be anticipated that the contact and response rate will be somewhat lower than that obtained in followups based on systematically established and maintained name and address files. Nevertheless, the information will be useful for obtaining initial estimates of levels and variations in the impact space and for obtaining information useful in planning later followups of initial input cohorts. This will also be generally true of the initial cross-sectional data, which will not be matchable and only be partially interfaceable across MISOE subsystems. Estimates of distribution parameters within subsystems can be established, as well as some time trend information. With this arrangement, it will also be possible to estimate correlations among variables between adjacent MISOE elements earlier than would be the case in a pure longitudinal design. Nevertheless, a sound dynamic simulation capability may be limited to relatively simple subsystems and associated flow models. Level and rate information will be available but only gradually will auxiliary modifiers of rates be systematically available from inter-element regression description.

The staff also proposes to replace an input cohort on each program only after a current cohort has completed the program rather than study every input group. This decision is logistically sensible, but it should be recognized that it may reduce the comparability of information across programs of different length.

Some Functional Issues in Data Processing and the Computer Facility

Two letters sent to the staff following the conference held in February, 1972 contain comments on Occasional Papers No. 1, 2, and 4, with a promise to comment on Occasional Paper No. 3. This section fulfills that promise in the present context of discussing general analytic considerations. Despite subsequent staff work on MISOE development, reflected in Occasional Papers No. 5 and 6, most of the comments in the two letters remain valid and reasonably consistent with material in this paper. Where exceptions occur, present commentary should override that in the letters. Nevertheless, those letters should be regarded as attachments to this paper as referenceable documentation in MISOE development.

Occasional Paper No. 3, labeled "very tentative", is addressed to the functional problems of handling information in connection with the computer facility. The organization of the data-entry subsystem is basically sound and includes provision for coding the stratification cell associated with a basic data record. Provision should be made in the data record layout for appending weighting factors, the number and nature of which will be more specifically defined in Occasional Paper No. 12.

The stratification cell code per se permits linkage only with the original sampling stratification cell weights. It is anticipated that one or more differential weights will be required for incomplete random sampling within schools and for correcting respondent data from followups for nonresponse bias. Not all of the weights will be needed in all analyses. Different kinds of analyses will require particular subsets of weighting factors. The weight actually applied to a particular record will typically be the product of individual weights in the subset. E.g., a student record* may contain three basic weights (possibly more), W_1, W_2, W_3. The weight applied in a particular analysis might be W_1W_2, W_1W_3, or $W_1W_2W_3$. As a minimum W_1, W_2, and W_3 should be on the record. It will be *convenient* for the *product weights* to be posted to individual records. Otherwise the product weight would have to be formed each time it was used in analysis and provision for this incorporated either in information retrieval operations or in computer program modifications.

---

*Ditto for a process records for a program within a school.

When a *cohort replacement* has occurred at the end of process, the input data for the new cohort is added to the system. However, the impact data for the original cohort may not yet be available so that retiring their records from the entry-level subsystem may be premature. Controlling product-impact analysis for input will require the capability of developing input-impact correlations as well as the previously computed input-process correlations, presumably retained in the analysis subsystem. Moreover, the development of weights to correct followup data for nonresponse bias will involve special analysis and comparison of the input characteristics of respondents and nonrespondents. Given adequate disc storage, entry-level data may be retained until no further use is expected; this solution hardly seems feasible in view of the expectation of conducting multiple followups over several years. The information can be removed from disc storage and stored on tape for reading into the computer processing unit, being sure that the complete ID, storage-retrieval, and linkage codes are retained in the tape record layout.

Provision must also be made not only *to printout* cumulations and averages, but also *to store* them for analysis, as indicated in Figures 1, 3, 4, and 5 of Occasional Paper No. 3. However, the sample data computations of basic statistics must consider the weighting factor (product of any relevant weights) as follows:

$$\begin{aligned} N &= \Sigma W \\ \Sigma X &= \Sigma WX \\ \Sigma X^2 &= \Sigma WX^2 \\ \Sigma XY &= \Sigma WXY \\ \overline{X} &= \frac{\Sigma WX}{\Sigma W} \end{aligned}$$

$$n\Sigma X^2 - (\Sigma X)^2 = \Sigma W \; \Sigma WX^2 - (\Sigma WX)^2$$

$$n\Sigma XY - \Sigma X\Sigma Y = \Sigma WXY - \Sigma WX\Sigma WY,$$

where $W = 1$ for census data and

the expressions on the left are population estimates from sample data. In general, W will not be a constant across summation and therefore cannot be applied externally to a summation, e.g., $W\Sigma X$. Anywhere the population parameters are known from census data, they should be used in preference to the sample estimators.

Most available analysis programs do not contain the weighting option. Some regression programs contain options for inputing either raw data or a previously computed correlation matrix. The necessity for adapting and modifying analysis programs, discussed in earlier sections, may be somewhat simplified with basic weighted accumulations held in hardware storage; nevertheless, having such capability in the programs as branching options will maximize flexibility.

Considerable attention is still required to the general issue of the degree to which level two analysis products should be kept in computer storage. Whole correlations matrices, some of large order may be developed for repeated but infrequent use. Moreover, the size of the matrices will grow as longitudinal data become available. One solution is to store these externally on tape and store regression results internally for easy call in dynamic simulation.*

The next part of this occasional paper discusses the MISOE repertoire of models and associated computer alogrithms. Therein will also be discussed the problem of variable selection and variable elimination, and the issue of relative importance of contributions that figures 6-8 in Occasional Paper No. 3 raise. The tentative formulation, not wholly satisfactory, served to lay these matters on the table and to indicate their general place in the MISOE development process. Comment on the optimization and simulation section of Occasional Paper No. 3 will be integrated with that on Occasional Paper No. 6, and deferred to Part Five of this paper.

*This requires careful tagging of this particular regression analysis from which such parameters were developed.

Part III. The Analysis Repertoire of MISOE: Models and Algorithms

This part of the paper continues the discussion of general analysis considerations across inquiries and MISOE subsystems with explicit attention to the repertoire of analytic models and their associated algorithms (what Occasional Paper No. 3 figures call "analysis options"). There is no need to elaborate further the discussion of the counting, aggregating, and distributing options, or to enter into extensive discussion of univariate distribution statistics. The system needs the capability of outputing weighted cross-tabulations, a capability provided by Data-Text or modifications thereof). Except for capability of computing phi coefficients and chi-square statistics (from weighted frequencies), the need for nonparametric statistics is judged to be low and therefore of low priority for MISOE. The ensuing discussion will therefore be focused on more complex analysis options. For each model discussed, the approach will be in terms of what the model does to accomplish what purposes, its relevance to, and therefore priority in, MISOE development and operations, and the inferential hazards in its application. It will be assumed that suitably weighted correlation matrices have been developed and stored under readily retrievable conditions for those models requiring them. Most general regression programs (e.g., BIOMED O2R) contain subroutines for selecting a subset of variables for a particular analysis (Cf. Boruch and Dutton's program VARELIM in Educational and Psychological Measurement, 1970, 30, 719-21.)

Multiple Linear Regression

With the possible exception of DYNAMO, the most general and most powerful analytic tool in the MISOE repertoire will be a highly general regression capability. The generality and power of the multiple linear regression model result from an appreciation of just what is "linear" about the model, and from the availability of computing algorithms which permit formulating a problem in a specified model .or developing the most efficient specified model for prediction from an a priori set of independent variables. Both sources of generality are important for MISOE. The general form of the model is: $\hat{Y} = \Sigma b_i X_i + C$.

The "linearity" of the model refers to the algebraic form of the equation in respect to the model parameters to be estimated (the regression weights, b_i), and not to the variables, X_i. In this context "linearity" has no reference to the shape of the scatterplots among the X_i or between X_i and Y. Therefore, X_i may be the value of any function of an observed variable or even a function of the other variables in the model e.g., the parabolic polynomial regression is linear in the above sense:

$$\hat{Y} = b_1 X_1^2 + b_2 X_1 X_2 + b_3 X_2^2 + b_4 X_1 + b_5 X_2 + C$$

Moreover, there is no restriction that X_i be continuous, quantitative variates; membership of an observation unit in a qualitative category (e.g., subject is male) may be indicated by dichotomous coding (1 if yes, male 0 if no, female; but recall the 2/1 transformation discussed earlier if zero is to be reserved for missing data). This notion also applies to canonical and discriminant regression (Tatsuoka). The power of the tool is further enhanced by the fact that a wide variety of hypotheses may be tested by comparing the R^2 for a full model with that for a reduced model consisting of an appropriate subset of the variables in the full model. A simple function of this difference is distributed as the F ratio familar in ANOVA. For elaboration and numerous examples of the power of this analytic tool, the staff should be familiar with Applied Multiple Linear Regression (Bottenberg and Ward, March 1963, PRL-TDR-63-6, 6570th Personnel Research Laboratory, Lackland AFB, Texas), and with Research Design in the Behavioral Sciences: Multiple Regression Approach (Kelly, Beggs, and McNeil, 1969, Southern Illinois University Press). One consequence of this for MISOE is that no separate programs are required for ANOVA and ANCOVA, which can be readily formulated in linear regression terms. Moreover, unlike classical ANOVA, the regression approach readily handles the "nonorthogonal case".

The other source of generality and power is the stepwise computing algorithm, which may be used not only when a set of predictors is specified, but may also be used to select the most predicting subset of variables for which data are available. Moreover, the variables are selected in order of their ability to add prediction of Y to that of the previously selected variables up to some stop criterion. This capability is especially useful for MISOE where the dependent variable may be a level variable in dynamic simulation; the selected prediction variables are then relevant candidates for other level variables in formulating the simulation model. Moreover, the associated regression weights, or functions thereof, are possible parameters in auxiliary equations modifying rates in dynamic simulation. The actual, not the predicted level of Y should be used in the simulation model.

One objection to the stepwise algorithm is that it tends to capitalize on sampling errors in the correlations. For this reason, and in the context of most anticipated MISOE applications, the stop criterion should be set in such a way as to reduce the number of variables entered while still giving a good approximation to maximum prediction. The BIOMED regression program controls on the F to enter or remove variables and can be chosen relative to sample size by correspondence with the associated probability level, with p of .05 entering more variables than p of .01. The computer printout shows the R^2 at each step and the variables selected at that stage. When, after several variables have entered, one is removed, a point has probably been reached where one is dealing with unstable artifacts based either on sampling error or the multicollinearity pattern of the system. One should stop iteration before that point. The program used by the Air Force Personnel Laboratory was based on the old Kelley-Salisbury technique of iterating on the regression weights, rather than on the n-th ordered partials. In that program, the stop criterion was an increase in the R^2 specified in the control card and for MISOE purposes would be set about .0004 (change in R of .02).

For exploratory purposes, more liberal stop criteria may be used than those suggested above, allowing more variables to enter. Also, it is not necessary to take the parameters of the *final* equation for use in subsequent analysis.

Greater flexibility in subsequent use of regression weights may be obtained by having both raw and standardized weights computed. Most programs compute and output one, but not the other. The subroutine to convert is simple to write and incorporate as a computer program modification. Which form to use when regression weights enter simulation equations will depend on the metric of the associated level variables. Some metric issues in dynamic simulation will be discussed briefly in Part Five.

The staff may find that some variables enter none of the regressions and if no other uses for such variables are found in MISOE operations, it may not be necessary to obtain data on them in the replacement cohorts. Conversely, if a variable enters rather consistently but weakly (late to enter; low regression weight) in many regressions, consideration should be given to obtaining purer measures of whatever factors are measured by the original variable.

The use of ipsative measures, at least in regression, should probably be avoided in MISOE. Examples are the Gordon and Edwards Personality scales and forced-choice interest instruments, which can be useful in guidance and counseling, but whose behavior in analysis is sometimes difficult to interpret. In some of these instruments, items scores are ipsative but derived scale scores nearly independent, in which case the misgivings expressed here are less relevant.

Relative Contributions in Regression

The results of a regression analysis express the pattern of functional relationships explaining observed differences in the values of the predictand, Y, in terms of observed or induced changes in the values of the predictors, X_i. It is beyond the scope of this paper to discuss the semantic morass and attendant philosophic issues implied by referring to such interpretation of regression results as "causal inference". Various analytic operations have been promulgated, however, for ascertaining the "relative importance", "relative contribution", "independent contribution", or "unique contribution" of the X_i to the prediction of Y. All of the methods depend on the pattern of correlations among the variables and none has any necessary reference to the temporal ordering of events presumed in "causation"; nevertheless, the practical importance of such operations is to give a partial answer to the question, "If I manipulate conditions such that the value of one X_i changes in the context of other X_j related to Y (which other X_j values may also change), what change in Y will result?" Answering this question is useful to a decision maker. Dynamic simulation adds the temporal dimension to this question and therefore to the way it is answered.

Operations for answering the question in regression terms are of three kinds: those primarily and directly depending on the rate of change in Y with respect to a change in X_i, as expressed in the regression weights; those accounting for variance in Y by partitioning partial regression variance and either implicitly or explicitly involving residual scores; and, a procedure for partitioning variance in Y in terms of the orthogonal factor variance of the system.

Interpreted as slopes, rather than as contributions to predicted variance in Y, regression weights are legitimate indicators of relative importance, and also of independent contributions of X_i in the special sense that intercorrelations among the X_i have been taken into account in the estimation of the weights. In another sense, they are not independent of each other since all b_i estimations depend on the X_j and the X_iX_j correlation pattern. The b_i lend themselves to formulating and solving dynamic simulation models, by their relative size indicating the need for incorporating the corresponding X_i probably as a level variable in an information loop, and its current mean as an initiating value. The b_i values may affect rates connecting levels in X_i and Y, either in rate equations or in auxiliary equations modifying rates; more likely the latter since b_i are rates of change of Y with respect to X_i rather than with respect to time. A word of caution: the b_i values are relative to the other variables included in the particular regression analysis and to the particular population or subpopulation on which they were estimated; presumably, then, the same variables should be involved as level variables and the same subpopulations should be involved in the dynamic simulation model in which they are used as moderators. Raw regression weights are also metric-sensitive.

The accounting for variance in Y in terms of X_i variances can also be used to facilitate choices of variables to include in a simulation model and the X_i variance contributions used in rate modifying equations. Moreover, the relative variance contributions, being ratios, are invariant under choice of raw vs. standard metrics, which is not true of regression weights. Partitioning of predicted variance also has the advantage that the contributions of process to product, or of process to impact, can take account of the influence of input variance, and should do so in dynamic simulation whether the input levels are explicitly part of the model or not. Thus, the variance contribution ratio of a process mix, used as a rate modifier in simulation, should not be contaminated with input variance.

Two frequently used procedures for variance partitioning may be summarily rejected unless the X_i are mutually orthogonal. One, based on the formula for the variance of a linear composite, defines the variance contribution of X_i as:

$$\frac{\sigma_i^2 + \sum^{j} r_{ij}\sigma_i\sigma_2}{\sigma_y^2}$$

The other, based on a formula from regression theory, defines the contribution as:

$$\frac{b_i r_{iy}}{\sigma_y^2}$$

Under conditions of orthogonality, the covariance terms vanish in the first procedure, and, in the second, the $b_i\ r_{iy}$ reduce to b_i^2 or r_{iy}^2. When, as usual, the X_i are intercorrelated, neither the σ_i^2, the $r_{ij}\sigma_i\sigma_j$, nor the $b_i\ r_{iy}$ terms are independent in any intelligible sense. Their only virtue is that they add up to the total composite variance.

Most other methods for partitioning regression composite variance depend implicity or explicitly on residual scores of the form, $Y_r = Y - \sum b_i X_i$. This includes procedures based on the Bottenberg-Ward comparison of two regression models, one being a subset of the other (e.g., the Creager-Valentine or Mood-Mayeske uniqueness-commonality model used in reanalyzing Coleman Report data). In these procedures, "independent contribution" means the amount of unique valid variance a subset of variables adds to the other variables (not in the subset) to yield the total variance in the full model. Although the variance partition of any one subset is orthogonal to the remaining subset, taken as a whole, the variance partitions among subsets (independent) are not orthogonal to each other; they add up to the total variance only because the commonality at the top of the hierarchy of "joint contributions" is estimated by subtraction from the total variance.

When the total regression composite has been built up by stepwise selection, the most valid and least correlated variables are likely to be picked, so that commonality partitions and such associated procedures as covariance control for input (discussed in the next section) may be reasonable and practical for MISOE. Even with some collinearity in the system, the inferential differences between partitioning variance in this way and using an orthogonalizing refinement should be negligible (see e.g., Creager, "Academic Achievement and Institutional Environments: Two Research Strategies", Journal of Experimental Education 40, No. 2, 197

The orthogonal partitioning of regression composite variance into the common and unique components defined by complete orthogonal factor analysis is applicable to any linear composite, including canonicals and discriminant functions (see Creager and Boruch, "Othogonal Analysis of Linear Composite

Variance", Proceedings, 77th Annual Convention, American Psychological Association, 1969; Creager, "A Fortran Program for the Analysis of Linear Composite Variance", Educational and Psychological Measurement, 31, No. 1, Spring, 1971; and Creager, "Orthogonal and Nonorthogonal Methods for Partitioning Regression Variance", American Educational Research Journal, 8, No. 4, November, 1971.) The advantages are that the partitioned components are all mutually orthogonal and additive to total variance, and can therefore be pooled across factors defined by subsets of variables of analytic concern (e.g., input, process, product or impact). To be maximally useful in practical application, the factors must be interpretable, either as a hierarchical structure like that of Schmid and Leiman (Psychometrika, 22, No. 1, March, 1957) or as an approximation to simple structure in which factor variance is spread across factors (normalized varimax rotation is popular). The procedure requires factor analysis of the regression system, and better factor definition can be obtained if additional marker variables are included in the factor analysis (their weights in the regression composite are zero and have no effect on the account of variance in that composite). Moreover, the factors may be less conceptually meaningful and communicable to MISOE users than the directly observed variables; in some applications, however, the factors can be regarded as more meaningful and the variables can be regarded as proxy measures of those factors. The delineation of space differentiations with their associated instrumentation classes suggests that such a view may already be implicit in staff thinking.

One implication of this line of thought is that some of the level variables in dynamic space may be factors rather than observed variables, but this would require the additional computation and storage of factor scores and development of trend information for use in defining rate equations for simulations. Whether or not this capability will be considered in further MISOE development, the variance accounting use of the orthogonal partitioning for input control (its original purpose) may be useful where the X_i are moderately intercorrelated.

The Control of Process-Product and Process-Impact Analyses for Differential Input

Serious inferential errors may result from conducting process-product and process-impact analyses without control for the nonrandom variation in input variables among students subjected to various process treatments. A given program or process variable may be unduly credited (or blamed) for changing students who differed before treatment or who would have changed the same amount an in the same direction given an alternative treatment (or no explicit treatment at all). Therefore, it is generally wise to pretest students at input time on product and impact variables and to use procedures which control analysis of process effects for input. An exception has been indicated which simplifies matters by making a plausible assumption: that input levels for objective skills in product space are constant at zero for secondary students in occupational education programs. While probably not strictly true, and ignoring possible differences on related student characteristics (e.g., psychomotor abilities), the assumption is probably a practical one. It is recommended, however, that such *not* be extended either to other variables in the product and impact spaces, or to other educational sectors (non-OE).

All but one method of controlling the process effects for input depend on some manipulation of residual scores, which may or may not be explicitly computed and manipulated. Astin ("The Methodology of Research on College Impact", *Sociology of Education*,1970, *43*) has reviewed several strategies using regression methods and which involve multiple-part and/or multiple-partial correlations and, in effect, are variations of multivariate analysis of covariance. These procedures are useful and practical. Critics point to the unreliability of residual scores, but usually support alternatives in which residuals are implicitly involved. Various schemes for stratifying, matching, and moderating have similar problems.

The main objection to such a practical approach as first regressing product on input and, then, regressing residualized product on residualized process, is that not all factors common between input and process should be treated as input (or as process, if the order of regressing the sets is reversed) sources of variance. They may be exogeneous or situational factors (e.g., cultural climate, community affluence) affecting both input and process, but not sensibly identified with either set of variables. E.g., the relation between family income and space per student might be a function of local community affluence, and if some student outcome (product or impact) were related to both, it would be quite doubtful whether management should be advised to raise salaries and wages in the community or raise taxes in the community to enlarge floor space at the local school; it would be quite dubious to partial family income out of space per student in the analysis.

With such a situation, by no means uncommon, the orthogonal analysis of the variance of a full model composite derived from free entry to both input and process variables may be helpful. Using a hierarchical factor analysis, the composite variance may come out on input factors, process factors, and on factors that are defined by a combination of input and process variables. With the latter explicit and interpretable, valuable clues about "reality" may result as well as some suggestions for formulating simulation models.

Principal Components and Factor Analysis

Most computer program packages contain routines for extracting eigenvalues and eigenvectors, and for performing factor analysis, including rotating transformations (often restricted to the normalized varimax rotation). Such capability will be required for MISOE if orthogonal analysis of composite variance, canonical, or discriminant analysis are anticipated. Moreover, they will be useful directly for descriptive analysis of MISOE data content within and across descriptive space subsystems, whether a factorial approach to simulation is contemplated or not. It will be useful to know the extent of within-space redundancy of information. The priority for such capability is moderate, being less than that for distribution and regression analysis, but otherwise valuable for MISOE to have in its analytic repertoire. Certain variations of factor analysis, such as alpha or image analysis, or the Guttman simplex,* circumplex and radex models are of doubtful utility for MISOE and such specialized capabilities may be deferred until the need for them becomes apparent.

The number of principal components or factors to rotate is popularly taken to be those with associated eigenvalues greater than unity. This practice has been challenged by Humphreys (Educational and Psychological Measurement, 24, 1964) when sample size is large, and by Shaycoft ("The Eigenvalue Myth and the Dimension-reduction Fallacy", mimeo available from the author), on both theoretical and empirical grounds. It is better to examine the plot of eigenvalue number against its size, looking for a break in the curve below a unit eigenvalue, but in any case to allow more degrees of freedom for rotation than permitted by the common rule. It will rarely be worthwhile to rotate vectors with eigenvalues less than .75, but one needs to retain after rotation only those factors which have loadings on more than one variable for defining common factor space. For the orthogonal analysis of composite variance, it is better to rotate too many than too few vectors because the purpose is to account for total variance rather than to minimize rank. If a good fit to hypothesized structure can be obtained by maximum likelihood methods, this capability may be useful (Joreskog and Gruvaeus, Educational Testing Service Research Bulletin, RB 67-21). Generally, oblique solutions will not be useful for MISOE except that the factors may more nearly match the meaning people associate with the factor name.

Canonical and Discriminant Analysis

An initial view of the multiple-space structure of MISOE suggests that a general capability to perform canonical regression relating one space mix to another would be indicated. With the exception of the special case of multigroup discriminant analysis, this is somewhat doubtful because:

---

*Not to be confused with the simplex algorithm of linear programming.

1. "mixes" will probably be better defined within spaces in accordance with simple regression against individual criteria to maintain flexibility and to relate spaces by simple regression of coded mixes, than to define mixes by weights that maximally correlate different space mixes. This is, however, debatable and further discussion of the point is invited (see also, Part Four below for further discussion);
2. canonical vectors may be difficult to interpret both within the staff and to external users of MISOE; and
3. the use of canonical information for dynamic simulation appears moot.

General canonical capability and associated models like Tucker's interbattery factor analysis and Hotelling's most predictable output mix are given low priority at this stage of MISOE development.

Multiple discriminant analysis has two parts: First, given an a priori set of groups to be discriminated (e.g., those "alumni" with certain product or impact mixes), define the discriminant space as a weighting of student characteristics (or process variables) which maximally discriminate the output groups (e.g., successful and satisfied, lawabiding and taxpaying citizens vs. welfare recipients, felons, and frustrated, angry protestors). The second part deals with the classification and allocation of personnel,* such as a new input cohort, to groups on the basis of their characteristics. Management not only has the option of changing process, but also of changing student inputs, by using guidance and counseling procedures which advise the student of likely outcomes of decisions (which should remain his) to enter certain programs or pursue certain occupational careers. Since the book by Rulon et. al., cited above thoroughly discusses this analytic area, further discussion here is unnecessary beyond a judgment that discriminant capability (and associated personnel classification and allocation) has a moderately high claim for priority in MISOE development and a strong chance that it may prove useful for certain classes of management decisions.

---

*This can be applied in an original data space as well as in discriminant space.

Temporal Analysis

The longitudinal aspects of MISOE and the dynamic simulation plans demand analytic capability in descriptive space for temporal analysis. None of the foregoing models and analytic capabilities pay any explicit attention to the passage of time; implicitly the information is temporally ordered by association with the temporal order of MISOE elements and by reference to cohort replacement sampling and followup measurement in impact space. This reflects staff recognition that it takes time for students to flow through the educational process and to "make their mark in the world", and that it takes time for management decisions to be implemented and for their effects to be felt.

The plotting of aggregate census data and of appropriately weighted sample data against time should provide some of the rate functions and some of the modifying auxiliary functions required for dynamic simulation, which is MISOE's major approach to coping with the temporal aspects of the state system of education. The problem for MISOE development is to ensure that variables whose values or distributions change over time with some naturally (i.e., without management decisions) and reasonably smooth frequency are repeatedly measured and that the information storage and retrieval system (especially coding) reflects the time of measurement, or more importantly, the time at which measured events occur. In the case of changes induced by management decision (whether under MISOE recommendations or not) change in level or distribution of a variable may be immediate, unique, and discontinuous, or may have scattered and delayed effects across the system. It would seem essential that MISOE have codable, storable, and retrievable knowledge of the nature and time of such decisions and of their implementation, if the system is to be able to reflect "reality" and if some dynamic simulation models will include information feedback loops. It is recommended that some attention to these issues and to their analytic consequences be given rather early in further MISOE development.

For the more natural and smoothly occuring changes, plotting routines are available in many program packages (BIOMED, DYNAMO), and may in fact be useful in describing more precipitous changes. Also, for the latter, Campbell's discontinuity regression concept may prove to be a useful tool, but it is not clear at present how this might be integrated with other analysis procedures for MISOE.

The whole methodology of lag correlations used in econometrics and in certain mathematical formulations of learning theory may be useful analytic tools for MISOE in dealing with temporal analysis. No attempt will be made to delineate these possibilities in this paper, but their potential utility for MISOE should have an early appraisal.

Miscellaneous Analysis Tools of Moderate to Lower Priority

This section deals briefly with some analytic tools of potential value to MISOE, but for which there does not seem to be any immediate demand for inclusion in system capabilities.

The first of these is path analysis, originally designed to delineate and investigate causal hypotheses in genetics, and in recent years, adapted and elaborated by sociologists. Typically, a hypothesized pattern of causal relationships is represented in a path diagram. The solution of a set of linear equations provides "path coefficients", which are usually regression weights or simple functions thereof, and which are interpreted to represent the strength of a particular path in the diagram. It is difficult to see what path analysis could do for MISOE, which is not better handled by the dynamic simulation model, where rates and their modifiers provide a more complete map of "reality" and time is explicitly taken into account. The only reference to time in path analysis is the use of arrows in the path diagram to connect level variables and to represent the temporal order of events. Staff familiarity with the logic of path analysis may be helpful in formulating dynamic simulation models. A recent paper gives a good introduction and list of references pertaining to path analysis (Anderson, James G. and Evans, Francis B., "Causal Models in Educational Research: Recursive Models", Working Paper No. 50, Institute for the Study of Social Change, Department of Sociology and Anthropology, Purdue University, 1972.)

Hierarchical grouping is an empirical taxonomic procedure of considerable potential value to MISOE. The need for it is not envisioned as imminent, hence the relatively low priority given here; nevertheless, the staff should consider adding such a capability to its analytic repertoire at a not too distant future date. It could be useful for defining student types, classes of process and product mixes, or "alumini" types (for subsequent discriminant analysis, as indicated above). The procedure is one way of reducing the large masses of data information in MISOE, where the loss of information on individual objects can be tolerated. Some key references to the logic and applicability of this model are:

Ward, Joe H., Jr. "Hierarchical Grouping to Optimize an Objective Function." Journal of American Statistical Association, 58, March, 1963.

Ward, Joe H., Jr. and Hook, Marion E. "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles." Educational and Psychological Measurement, 23, No. 1, Spring, 1963.

Bottenberg, R. A., and Christal, R. E. "An Iterative Technique for Clustering Criteria which Retains Optimum Predictive Efficiency." WADD-TN-61-30, Personnel Laboratory, Lackland AFB, Texas, March, 1961. (Clustering of regression equations in terms of homogeneity of regression).

Rock et al., (American Educational Research Journal, Vol. 9, No. 1, Winter, 1972) have proposed and illustrated a strategy for studying process effects by grouping programs on the product-input regressions and then using process variables to discriminate the groups equated for input. The strategy combines regression, hierarchical grouping, and discriminant analysis, and is too new to permit a fair appraisal. One difficulty that may be encountered is that the regression composites within some of the smaller programs may be too unstable to carry the grouping and discrimination load which follows.

The "policy capturing" model is a special case of static simulation using regression analysis to simulate subjective (or aesthetic) human judgements. This is done in terms of the objective information available to the judge(s) about the set of objects being rated or ranked. It has some interesting possibilities for MISOE and for the management of MISOE in the latter's interaction with representatives of societal space. It is quite conceivable that MISOE might want to simulate (dynamically) for state level management the effects of local policy judgements on the state system, where such judgements are made on the basis of subjective weighting of information available either locally or through state and regional communication channels. This may be useful in incorporating and using information feedback loops in dynamic simulation models. One outcome or policy option for the state level manager may be selective emphasis in information dissemination; alternatively, the MISOE management may want to know the relative weights that educational management gives to information in the system, whether or not that information is MISOE input or output.

To apply the model requires collection of ratings or other scaled judgements on a set of objects (programs, allocations, relative importance of societal goals, etc.), measures of the information available to the judges (a single manager, managers, a committee or panel) about the objects rated, and the regression package. The multiple R is usually very high and measures the validity of the policy capturing simulation, and the weights give the substantive information as averaged across the judgements. The technique is also useful in aiding panel consensus by feedback of its results to the judges. Two papers by Dr. Raymond E. Christal, Personnel Laboratory, Lackland AFB, Texas are relevant:

> "Selecting a Harem -- And Other Applications of the Policy-Capturing Model", PRL-TR-67, and
>
> "JAN: A Technique for Analyzing Group Judgement", PRL-TDR-63-3.

The latter involves integration with hierarchical grouping.

The National Academy of Sciences - National Research Council uses the procedure in the evaluation and selection of candidates for National Science Foundation graduate fellowships. The multiple correlation between objective and codable information provided to the judging panels and the judged ratings of applicants has consistently been about .85 over several years. This is only slightly less than the estimated reliability of the panel judgements.

Part IV. Noneconomic Analysis Considerations Within and Among Subsystems of Static Space

Introduction

Although Parts II and III discussed analytic issues of concern regardless of the MISOE subsystems involved, considerable reference was made to these subsystems. Nevertheless, more specific issues are raised in Occasional Paper No. 5 for descriptive analysis in static space, and in Occasional Paper No. 6 for simulative analysis in dynamic space. This part focuses on the more specific analysis issues raised in static space; the next part focuses on those raised in dynamic space. Primary concern in the subsequent sections of this part will be with educational space and with the educational post-impact space (see Figure 1 of Occasional Paper No. 5). However, the need to define optimal process and product mixes by student type, combined with the fact that available data represent the *status quo*, poses a special problem for analysis not considered in Part Two above. The next section, therefore, discusses this problem, prior to giving specific attention to the analysis problems within and among the process, product, and impact spaces.

The Range Restriction Problem

Much of the initial data for MISOE come from the present structure and "student flow" characteristics of the operating educational system. As noted in Part III, above, prior educational management decisions about what kinds of students enter what kinds of programs (with their associated processes, products, and impacts) results in differential inputs to the various "pipelines". Thus all data about students within any process-product-impact channel reflect the status quo. Managers will want to know the result if some student mixes not presently in a channel were permitted or encouraged to enter this channel, or, put another way, the result if a given student mix were to go through a different educational channel. The earlier discussion of controlling analysis for differential input was concerned with reducing the risk of inferential error when comparing results of analyses across channels (e.g., OE vs. general vs. academic; TV repair vs. automechanics; automechanics is school A vs. that in school B), or when judging the efficacy of a process within a channel. In the search for optimal matches between student and program mixes, or the search for optimal mixes within spaces given a fixed mix in another space, data for currently nonexisting matches of students and programs will not be available for comparison. This implies that optima may be missed. Moreover, inferences from analyses carried out within an IPPI channel will be strictly relevant to the status quo for that channel.

In dynamic simulation, initial values of level variables can, and often should, represent the status quo, and then, additional simulation runs can be made with different values specified by hypothesis. However, if some of the rates and auxiliary modifying equations are to express relationships derived by within-channel regression analysis, an assumption is being made that these relationships will hold for alternative student input mixes. The assumption may well be false and therefore inductive of inferential errors.

What is involved analytically is that the correlations based on a particular student input mix (whether or not the correlations involve the student characteristics variables) will generally be smaller than those based on the entire student input population, or on subsets of that population that include alternative student inputs to a special channel that might be under consideration (e.g., an LEA concerned with students in the local community rather than with the whole state distributions). Moreover, the attenuation of correlations from "restriction of range" along one or more dimensions of student space will be nonuniform across a set of variables, thus distorting the pattern of correlations in a matrix in addition to lowering their average value. This situation distorts all channel regressions and regression parameters and distorts the regression techniques for controlling channel regressions for differential input.

Two kinds of formulas exist for "correcting observed correlations for range restriction". One is applicable in some MISOE situations to correlations between student variables and process, product, or impact variables; the other in some situations to correlations among process, product, or impact variables the variance of which has been restricted by student input selection. Each of these formulas exist for correcting single correlations for selection on a single variable, and in their multivariate generalizations, permit correction of whole correlation matrices for restriction on one or more correlated student variables.

These formulas are presented with references to books by Thorndike and by Gulliksen in a set of memoranda attached to this paper as an appendix. It is recommended that uncorrected correlations be stored and retrieved, and if correction is required for a particular analysis, it can be done prior to entering the correlations into regression analysis. The correction requires an *ad hoc* Fortran program, probably not available in commonly used packages.

Awareness of the assumptions about the nature of the restriction on which these formulas are based may guide staff judgements about their use in MISOE. Basically the formulas assume that:

1. restriction was caused by truncation (e.g., applying a cutoff score for admission to a channel) on a variable,
2. this truncation was strictly adhered to,
3. the raw score slopes (b_i, not betas) of the regressions involved were unaffected, and
4. the variables were all measured without error.

Regarding the last, the correction procedure could be entered with correlations corrected for attenuation due to measurement error. The first three assumptions are plausible in the military situation for recruits assigned to training programs (the situation for which the formulas were developed and most frequently used), but not always valid even then. They are plausible in local situations where the educational manager has specified and enforced such cutoffs in selection (e.g., a minimum IQ to enter this program), and where the standard deviations are known for the sector of student space served by his jurisdiction. Even under ideal conditions where the b_i are unaffected, one may prefer to use "variance accounted for" in interpretation and simulation, for reasons discussed in Part III.

More serious are those MISOE situations for which these formulas would be of limited or even dubious value, but where the problem remains. In one example presented by the staff, the principal interviews the "applicants" for a program in his school, and judges the "interest and motivation" of the student for that program. In this case, one could probably use the formulas (assuming one has an external measurement of the relevant interest) even though the exact cutoff and the consistency of its application may be unknown; the formulas require measurement of the effect in terms of a comparison of standard deviations. Another example is the situation where analysis is being performed on information pooled across schools giving the "same" program, but with variations across schools in actual admissions criteria applied to various pools of potential students. Partial solutions for such situations are presented in a memorandum on "simulating" complex selection, appearing as an appendix to this paper. Under some conditions it is even possible to regenerate a normal bivariate scatterplot seriously mutilated by complex selection realities, by iterative operations based on the discrepancies between pre- and post-mutilated marginal distributions. This possibility is described in another memorandum attached to this paper. The efficacy of these suggestions in a practical setting is unknown.

For some analyses involving product and/or impact correlations (within and between spaces), the range restriction corrections may be involved where allowance must be made for selective losses due to dropouts. Moreover, for MISOE to make the kinds of comparisons across occupational, general, and academic educationa programs in terms of general educational development requires not only input GED measurement in student space, but also allowance for the multivariate restriction of range implied by differential selection on achievement among these "tracks", regardless whether the choices are made "freely" by students or as a result of some kind and degree of management intervention.

Analysis Considerations for Noneconomic Factors in the Process Space

Staff delineation of the process space is documented in Occasional Papers No. 2 and 4, with an addendum to the latter included in Occasional Paper No. 6. Although some comments on the contributions of papers 2 and 4 were included in the two letters sent following the February conference, this section develops some issues raised there and in Occasional Paper No. 6. In several ways the process space is the heart of the system and a major source of its complexity. It is also the major MISOE subsystem in which economic and noneconomic aspects interface with each other and with implications for developing rate and auxiliary modifying functions in dynamic simulation. Proper treatment of this important topic will require later integration of the concepts currently being developed in Occasional Papers No. 7-12.

The process space involves both description of the climate of learning in terms of human, physical, and organizational factors (see Occasional Paper No. 4), and that of the content and sequencing of instructional events (units) as organized into blocks for each program. Analytic capability must be provided for both kinds of process descriptions and their interactions in the vector formulation of a process mix. The vector, or subvectors (formed from a subset of the defining variables) can, of course, be coded so that student types can be matched with process and product mixes. Process information should be obtained, in accordance with the sampling design, locally, within a program within a school, so that pooling of data on common variables can be accomplished across schools and programs, locally and at regional and state levels. "Interactive" process variables may either be directly observed, e.g., the number of students on a piece of equipment, a human-physical combination, or be generated as needed in the form of $X_i X_j$ terms, e.g., the joint occurrence of a teacher characteristic (more experience) and assignment to a physical factor (the better equipped of two available laboratories). The latter example is likely to occur, but for a better product over all students in that program at that school, the more experienced teacher may be better able to adapt to less ideal physical arrangements at the same salary and equipment cost.

In a particular program there may be variations in content and sequencing of instructional events from one school to another (or one locale to another). Content variations (additions or deletions of particular units or blocks) permit investigation of their efficacy in terms of products and impacts. If something like 85% of the schools giving a program have the same content structure (blocks and units) and 15% have one or more variations, a dichotomous code can be defined permitting regression comparisons over schools. If a program has something like 50% common structure across schools, it is likely that additional dichotomous variables can be defined to tag additional variations across schools in the content structure of a program. Sequencing of blocks, or of units within blocks, within a program, can be similarly treated. Precisely what is feasible will depend on the counts of such content and sequencing variations. Greater flexibility may be obtained if the process record shows an actual sequence, e.g.,

$$\overline{\underline{/\ 2/\ 1S/\ 4/\ 3/}}$$

could indicate that unit 2 is given first, units 1 and 4 are given concurrently

(S denotes unit given simultaneously with the next unit shown), followed by unit 3. Where this is variable for students rotating sequences to maximally utilize available equipment, such a sequence code should be posted to the student record along with data that indicate that the particular student went to a particular school and took a particular program. (Note: this kind of cross-linkage between student and program information is crucial for MISOE; some of it is a matter of appropriate codes being placed in information records, a format or layout problem; some of it is a matter of the addressing in the information storage and retrieval system. The staff appears to have awareness of this and to be making appropriate provisions).

Similar logic and treatments are relevant to variations among students and schools in time-spent-per-unit. The example on page 46 of Occasional Paper No. 5 in which students move on to the next unit regardless of performance may not always be the case for all programs, schools, and levels (and in any case, a relevant question is whether the policy is a wise one). Many of the variables descriptive of the general setting and specific instructional climate may be indicated by dichotomous coding, permitting easy generation of codes for joint occurrence of process characteristics.

Some of the above suggestions imply long data records for the process space. They may also imply the need to expand the information storage and retrieval system as delineated in Occasional Paper No. 4, figures 1 and 2. Occasional Paper No. 5 Pages 45 - 49 indicates staff progress in keeping this system operationally flexible. It may be necessary for the staff to prepare a document fully delineating this system in the light of Occasional Papers No. 7-12.

The notion of using a standard form for collecting process information for each program is an excellent one, and the indication on the form of the storage-retrieval(or other)codes should facilitate carrying document information into computer storage. The idea of assigning a process mix number is also useful, but it should be noted that a total process mix may contain subvectors or submixes of frequent and selective interest. The only question is how far to carry this. Perhaps one submix code would be for the human-physical-organization factors, or one code for each of its three component submixes, and another would be for the content and sequencing information. Each submix, however defined, should have its own ID number.

Units of analysis conducted within process space are likely to be schools having a given program. In some analyses comparisons of selected climate information across programs within a large school may be desired. This is in contrast to student space where students are likely to be the unit of analysis, and in product and impact spaces where either students, schools, or programs may be the analysis units. A given analysis across MISOE subsystems will have to deal with this. A particular inquiry will have to be judged as primarily focused on answering questions about what happens to students or in terms of what processes are under study. In the first case, students and/or student types are followed through the system and for each the appropriate process mix is retrieved and merged with student input and output data. In the second case, the data for students entering and leaving a process are averaged over those entering a particular process mix (in terms of schools, programs, e.g.,), retrieved and merged with similar averaged output information. One will usually obtain higher multiple correlations in the second case, but the real question is which is appropriate for a given inquiry (or subinquiry): are we concerned with what processes do to individuals, or, with what processes do, in the aggregate, for society? Overall, both, but not within the same specific analysis. The same question applies in dynamic space and must be answered the same way where regression information is to interface and provide input information or simulation. Even if they can be mixed in simulation, separate regressions by units of analysis will be required.

Occasional Paper No. 5, Page 48, proposes to post the weighting of each process variable in the mix in which it occurs and the average weight of a variable over all mixes as part of the storage and retrieval of process mix data. This hardly seems realistic as a solution to a real problem. In general, it is inconsistent with the flexibility requirement. More specifically, it ignores the multiplicity of weights a given variable can receive within a single mix depending on the regression analysis in which a weight (b_i or partitioned variance) was estimated. The same variable may have quite different weights for various product and impact mixes and submixes, over different analysis units and aggregations thereof. The problem would be compounded when trying to average the weights a variable receives across process mixes; this is a dubious practice anyway, instead of recomputing them from aggregate correlation matrices.

There may be much more homogeneity of regression in the system than the above criticism assumes, but we don't know this. The homogeneity issue can be answered by special application of the Bottenberg-Ward procedure or by hierarchical grouping of regression equations. It is also possible to group or cluster mixes within process space without reference to IPPI relations.

This will not weight the process variables, but merely classify mixes without direct interfaceability with other MISOE components. It may, however, be useful for organizing a listing of mixes with or without their associated cost and weight information. It may require a separate information storage and retrieval section with separate but cross-linked addresses for storing the costing and weighting information.

Analysis Involving the Product Space

From an analytic viewpoint, the product space, like Janus, faces in two opposite directions. Students within programs have been "processed" and come off of the pipeline as "program completors". Data about them may constitute the initial set of dependent variables in the analysis of process; impact variables constitute a later set. But the product variables indicate the educational managers' assumption that product quality is related to impact on societally defined goals. This assumption is validated using product variables as independent variables in the prediction of impact, i.e., product-impact analysis. The two major purposes of product data within MISOE as delineated on Page 79 of Occasional Paper No. 5 will be served primarily by process-product analysis. However, both need input controls ("by student type" and/or variance controls for differential input), and the second is further served by process-impact analysis.

Occasional Paper No. 5 distinguishes gross and specific types of product data. Gross data, such as the number of students completing a program would be obtained for all educational sectors (OE, Academic, and general). Similar gross data should also be obtained for each program about the number of dropouts. Moreover, the student records should clearly indicate for each student not only the program entered, but his completion-dropout status. Some so-called "dropouts" may actually be transfers to another program or even to another sector, where they may or may not become "completors". The data system should be able to reflect this reality.

Occasional Paper No. 5 indicates specific product data will be obtained for completors of occupational programs in cognitive, psychomotor, and affective (primary attitudinal) areas. Consideration should also be given to obtaining some affective data on dropouts and transfers since this may relate to later employment status and other impact data. Similar reasoning suggests obtaining some affective data in the non-OE sectors in addition to the GED information already planned.

This kind of thinking implies legitimately missing data for dropouts on some of the process and product data, because such data are not applicable. The earlier discussion of missing data refers to data that should be present and usually is, but is not obtained for some observation units. In the present case, where process-product data are missing due to noncompletion of a program, no replacement values should be computed. Care must be taken in regression and other correlational analysis of process-product, process-impact, or product-impact relations to perform the analyses on completors only, dropouts only, or if across all inputs, to use process and product variables obtained on both completors and noncompletors

Rational management decisions about which students should enter which programs cannot rely solely on gross counts of completors by student type. The specific product data, by student type, is also relevant to this kind of a decision. It is the kind of question that requires the combined application of taxonomic and discrimination models. One could classify students into types within student space and carry student mixes through process-product-impact analyses. The taxonomic nuclei could be defined on a random or self-weighted sample of the students in the general sample space, assigning all other students by the personnel assignment algorithms. It would probably be better, however, to define output groups based on product and impact data, and to use student data to classify students in terms of such output groups, weighting the student data to maximally discriminate the groups and "assigning" new student cohorts to programs in which they will have maximum likelihood of achieving high product scores and be most likely to have favorable societal impacts.

Occasional Paper No. 5 notes that the educational process must be flexible so that improvement is possible, and that product objectives must not be overprescribed. This is not envisioned to encourage vagueness in specifying objectives, but to allow objectives to be added or deleted in a program over time, and to allow variations across schools in specifying objectives. This implies the same kind of analytic flexibility in product space as was discussed earlier for process space. Provision must be made not only to obtain product data on unique objectives but to ensure storage and retrieval linkage between unique processes and products. Moreover, the fact that a product datum refers to a unique objective and the process for achieving it needs to be so tagged to ascertain in product-impact analysis whatever unique contributions to impact such unique products may have. Although primary reference here is specific to product data related to a unique objective, the gross counts of numbers of students completing a unique objective should be obtained. Moreover, it is well to keep in mind, for both common and unique objectives, the possibility that the process associated with a particular objective may affect other products in addition to the one for which it was promulgated.

In some programs uniqueness may be introduced in schools not in the sample. As soon as possible after this occurs, consideration should be given to including the school in the sample at the next cohort replacement time for the program. Some weight adjustments are implied and the feasibility of this notion depends on the frequency of the occurrence of unique changes, presumably knowable from census data on programs.

Each of the three specific data types: cognitive, affective, and psychomotor performance have some measurement and analysis implications. As one possibility for system expansion, it might be useful to obtain product data not only from program staff and program completors, but also from employers or supervisors of those students who were on work-study process plans. Such data might be obtained in the form of a set of rating scales on all three performance areas, and might well have some predictive validity for certain impact measurements (e.g., employed vs. unemployed; on-the-job performance ratings in impact space).

Cognitive measurements by pencil and paper tests will result in distributions of test scores. Conversion to pass-fail in terms of some cutoff specified by a cognitive objective loses considerable and useful analytic information. It will contribute to flexibility to have both the "continuous" and dichotomized test scores available: the former for regression analyses across spaces, the latter as an elaboration of the gross accountability data.

It may be instructive to note how the comparison of general educational development across sectors might be formulated in regression terms. The "full" model is defined as:

Y = GED product score, the dependent variable
X_1 = Input GED score for input control
X_2 = Dichotomous score for taking Academic tract
X_3 = Dichotomous score for taking General tract
X_4 = Dichotomous score for taking OE tract; program 1, mix 1
X_5 = Dichotomous score; program 1, mix 2
X_6 = Dichotomous score; program 2, mix 1
X_7 = Dichotomous score; program 2, mix 2

$$\left.\begin{aligned} X_8 &= X_1 X_2 \\ X_9 &= X_1 X_3 \\ X_{10} &= X_1 X_4 \\ X_{11} &= X_1 X_5 \\ X_{12} &= X_1 X_6 \\ X_{13} &= X_1 X_7 \end{aligned}\right\} \text{Input scores by tracts}$$

It can be simplified or expanded by the gross vs. fine attention to programs and mixes within occupational education. The first seven vectors might be expected to be retrievable from observed and stored information. The first vector, from student space, is included primarily to permit the generation of product vectors, X_{9-13}. Even without such product vectors, the inclusion of X_1 in both full and reduced models will give an overall control of a test of some hypothesis involving the other vectors. Vectors X_{2-7} indicate which students went through which educational channel. The zero-order validity coefficients (r_{xy}) for Vectors X_{2-7}

correspond to uncontrolled t-tests for contrasting a particular channel against all others with respect to the GED product score. Reduced models containing two or more of these vectors permit tests of contrasts between the pooled channels retained and those dropped from the full model. Thus, all the OE channels can be pooled and contrasted with non-OE channels. The vectors, X_{8-13}, permit tests of homogeneity of regression among two or more channels. This rather special and simplified example illustrates the power of the regression approach to handle whole series of ANOVA, ANCOVA, and regression homogeneity problems from one formulated "full" model. The references cited in Part Three must be consulted for further details and for a wider range of examples of the power of the regression model.

The affective data consist primarily of Likert and Semantic Differential scales of attitudes toward self and work. Retests of some personality measurements (e.g., authoritarianism)pretested in student space may also be helpful, since personality changes, whether or not attributable to the educational experience, may be predictive of impact variables. The proposal on Page 88 of Occasional Paper No. 5 to treat affective data separately from psychomotor and cognitive data is reasonable when product variables are to be dependent variables in process validation. When, however, they are used to predict impact variables, it is quite feasible to combine the three types of product data in analysis and this permits the examination of possible interactions among product data types in predicting impacts. Affective data on dropouts and transfers may also be important information to obtain.

Affective objectives (and therefore data) reference programs, but not blocks and units. Stipulation of the objectives "within department faculties" implies possible variations across schools with the kind of uniqueness problems discussed earlier (with similar treatment recommendations). Input control is just as important in analysis of affective data - perhaps more so - as in analysis of cognitive data. The discussion of Figure 9 in Occasional Paper No. 5 (Page 90) ignores differential input and leads to an inference that school A fosters better attitudes than school B, which may be true, but may also be an inferential error.

The performance objectives, largely psychomotor in conception and measurement, probably involve cognitive and affective components. The use of pass-fail measurement of achievement of specific objectives in this domain is quite reasonable. The objectives are like items on a test with item scores determined by one or more scorers (raters) observing performance, either directly or by reference to video tapes. The inter-rater reliability of determining pass-fail on a specific objective is a function of the number of raters; that of a single rater might be as low as .30. The Spearman-Brown formula can be used to estimate the reliability of pooled judgements for a given number of raters. If a single rater reliability is .30, that for two independent raters is about .46; for three, about .56, and for four, about .63. The use of video tape and discussion between discrepant raters should improve the reliability of the ratings (not necessarily their validity) and hence, permit the use of fewer judges. Inter-rater reliability can be used to correct correlations of ratings with other variables for attenuation from measurement error.

If the objectives are scalable, the scale scores can be used in analysis. In the case of Guttman scaling, reliability of the scaled scores comes out of the scaling process itself. For a set of objectives to form a true Guttman scale, they must form a unit-rank correlation matrix, i.e., reference a single common factor. This univocal feature of such a scale provides a score with a very narrow band-width with the usual advantage of a clear meaning of what is measured, but the disadvantage that the scale will have slim chances of correlating very highly with external variables. For this reason, one doesn't often hear of the extensive development and use of Guttman scales in large-scale practical programs. Nevertheless, this approach is well worth trying for MISOE. It is more likely that performance objectives within programs will form one or more "quasi-scales" in the Guttman sense, and these should be quite useful if reliabilities can be maintained in the .75-.95 range. It may be helpful to generate a matrix of phi coefficients (phi/phi_{max} if objectives vary considerably in difficulty, i.e., percent passing), and perform an informal clustering of performance objectives to ascertain which sets are likely to scale.

The staff recognizes the fact that some objectives will not scale and proposes a procedure (Figure 7, Occasional Paper No. 5) for assigning unique numbers to patterns of achieved objectives. This can be done whether the objectives are scalable or not. The procedure ensures a unique number will be assigned to each possible "mix". These numbers, like those on the jerseys of football players are nominal, they tag the patterns, but do not scale them. The pattern numbers should not be used analytically, but the presence or absence of each pattern indicated for each student as a dichotomous variable. The pattern number could be what is stored so that the dichotomous vectors can be readily generated as needed for analysis.

The utility of product data to management is defined on Page 79 of Occasional Paper No. 5 in terms of maximum product for given cost and/or least cost process to achieve specified products. These questions involve the integration of economic analysis with noneconomic interspace analysis, an integration possible when papers 7-12 have been completed. Dynamic simulation should be useful for resolving management alternatives, given status quo simulation followed by runs in which product levels and costs are changed in search of optimizing combinations.

Analysis Involving the Impact Space

The variables of the impact space are to indicate societal values, societal action goals to realize those values, and to constitute the ultimate criterion space for management policy evaluation and decision making. Although aggregations of these data over educational and noneducational sectors, and over schools and programs, are of direct importance, the actual impact data for each individual will be needed and identified as such for static interspace analyses. Aggregated impacts will be of maximum interest to legislators and state level managers. Interactions among impacts (mutually enhancing or constraining them are virtually ignored in present planning but may be of interest as MISOE expands. Aggregated impact information will probably be critical level variables in dynamic simulation and, indeed, certain kinds of dynamic interactions among impact levels can be hypothesized and included in formulating dynamic space flow diagrams.

In a narrow sense, impact space measures the benefits in cost-benefit relations. More broadly, impact is often economic, too, in that there will be interest in economic benefits, both societal and personal. A further delineation of this view is part of the anticipated integration of economic and noneconomic considerations.

At the stage of formulation of management inquiry and translation into analytic operations, direct interaction between representatives of legislators or managers, and MISOE personnel is anticipated to be necessary. Even with extensive education of inquirers by MISOE staff over a period of time, it is unlikely that interrogation of the system can be confined to inquirer manipulation of remote computer terminals. Certainly that is a useful part of MISOE: but the computer cannot generate the interim decisions between problem formulation and analysis; choice of relevant data, selection and ordered application of appropriate models, algorithms, and interpretation of analytic results. The computer will neither formulate the flow models nor write the model equations for dynamic simulation.

Much of the initial interaction between MISOE staff and inquirers will involve efforts to get exact specification of the problem at hand in terms of:

1. level of application (i.e., subpopulation referenced by the inquiry),
2. what is to be optimized, or other goal of the inquiry,
3. what is an acceptable solution,

4. the time by which a goal is to be achieved, and
5. in the case of multiple, related goals, what priorities are assigned to their achievement.

The discussion of impact space in Occasional Paper No. 5 shows staff awareness of these issues.

The possibility that relative priorities for multiple goals may be under consideration implies for static space analysis that weights (probably ratings or rankings) of the relative importance of impact goals be appendable to impact data. When predicting a single impact, they will not be needed, but in formulating and predicting an impact mix, such capability should be selectively and flexibly available. Note that different inquirers may have different priorities. This means that priority weights should not be appended to data within the information storage and retrieval system, but be flexibly introducible cojointly with retrieved data in applying an analysis model. Note, too, that such weights are in addition to any sampling weights which must be appended to the data.

It is recognized that impact data may come from several sources including gross summary data from other agencies,and more specific followup of "alumni" from academic, general, and occupational education channels, including dropouts. In addition to employment and citizenship information, it would be desirable to ascertain the location and mobility of former students,both substantively and as an aid to updating name and address files. Impact mixes can be formulated as person vectors on impact scores, and at aggregate levels, mean vectors appended to aggregate vectors of data from other sources. These two kinds of impact mixes that can be combined at more aggregated levels correspond to the direct-personal vs. indirect-societal dimension of impact classification shown in Figure 2 of Occasional Paper No. 5. The immediate impacts are more closely associated with process and product data and to the transition between pre- and post-impact distributions, but can be treated analytically in the same way as the longer range impacts developed from followup data.

From an analytic viewpoint the process of setting societal and related impact goals is "a given", underlying the problem for analysis. The process of goal setting is described for analytic purposes quite adequately in Occasional Paper No. 5, at least for anticipated impacts. Insofar as unanticipated impacts (i.e., those not related to specific educational goals) are in fact anticipated as possible conditions for former students, and are measured, they pose no serious problem for static analysis. Their levels, however, may be important to include in dynamic simulation models.

The educational pre-impact space is defined as that for storage of "existing" levels and rates (i.e., ratios, not necessarily rates in the dynamic simulation sense) of impact variables. The "initial data" collected on impact variables from previously "processed" students and from other sources can be stored here. As a cohort in each program moves out of educational space, their impact data and concurrently updated data from other sources can be placed in post-impact space. As the next cohort moves out of educational space, impact data on the previous cohort can be moved to pre-impact space as "existing" data, replacing the "initial data" and making room for the post-impact data on the new cohort. This will automatically preserve and update the distinction between pre- and post-impact spaces as MISOE operates over time. It may be desirable to retain cohort data retired from pre-input space externally on tape for archival reference, if necessary. In any case, the impact data time and cohort must be clearly indicated in the information storage and retrieval system.

Many analytic considerations involving impact data have been discussed in earlier sections of this paper, because analysis of other spaces will often involve the impact data. There remains the issue that certain post-impacts will be credited to (or blamed on) educational processes and products, even with input controls, where the observed impact may be the result of ongoing cultural and economic processes. The effects could occur between program completion and long-range impact measurement, and differentially by the locale in which a completor lives. No attempt will be made here to cope with the potential risk of inferential error so introduced, nor should the staff be unduly concerned (although aware) with coping with it in early development and implementation of MISOE. Changes observed in the noneducational control group will provide some clues to the nature and extent of the problem and expanded MISOE can be designed to cope with it, if necessary..

The Educational Human Input and Student Spaces

The staff has deemed it convenient to distinguish a societal resources space from the original input space, now called educational resources space (Figure 1, Occasional Paper No. 5). Each is subdivided into the human and economic components: the educational human input space and the student space, respectively. Since essentially the same variables and observation units (those who become students) give rise to similar aggregations and mix formulations, the two spaces may be considered as one for most analytic purposes. Nevertheless, the interest of state level managers will focus on the educational human inputs, implying emphasis therein on the aggregated data over all educational sectors, schools, and programs, and over certain demographically defined submixes. Initial emphasis will be on status quo distributions and a priori grouping of mixes. One may also anticipate that some demographic, ability, and achievement measurement at late primary school levels could be involved. If such a distinction between educational human input and student input information is contemplated, the information storage and retrieval system should reflect it along with the capability of tagging at the individual level who enters which educational channel.

The student input information is required to characterize the sorting out of students into channels, whether this tracking is accomplished via student choice, via administrative control, or a combination of the two. As part of the process data, any cutoffs for entry to a program should be posted for possible use as constants in rate equations in dynamic simulation. The student input information is also required to control analysis validity for differential input as discussed earlier. An analytical question arises with regard to input control through methods discussed in Part Three vs. conducting analysis within "student types". The more finely defined the student types, the less need there will be for the variance controls on input, but also the smaller the number of observation units on which the analysis can be based. Analysis within moderately gross classifications of students should use the variance controls for within-group heterogeneity.

The full pattern of characteristics and descriptions for each student, including his educational channel, constitutes a total mix from which a variety of submixes should be flexibly derivable for different purposes. For purposes of some state level policy makers, mixes of demographic, ability, and achievement measures may often be all that is needed; larger submixes may be needed (including personality data) for regional policy makers, "program directors", and for analytical controls. Aggregation and classification of mixes for state level analysis may reasonably be rather gross and may involve categories along continuous dimensions (e.g., high, medium, low IQ). In the early implementation of MISOE the classification of "student types" for educational human input may be a priori, to be replaced by a more sophisticated taxonomy based on IPPI relationships to be developed as early cohorts go through.

The taxonomy of student space involves so many students and so many variables that some a priori grouping in terms of educational sectors may be helpful. If hierarchical grouping is used to define the types, the grouping should be based on a distance matrix with the objective function of maximizing the ratio of the among-groups sums of squares of distances to the within-groups sums of squares of distances. The grouping is apparently insensitive to whether D or D^2 is used, or whether distances are computed on the score vectors (Cronbach's D^2) or on their principal components (Mahalanobis' D^2).

Analysis Across IPPI Spaces

Most analyses of any substantive value beyond distributional description and aggregations on particular variables and mixes will be across the IPPI spaces. Because so many variables are involved for which interaction vectors may be meaningfully generated, regression analysis across spaces should first be performed so that the stepwise algorithm can reduce the number of relevant variables. Then, plausible interaction vectors can be generated involving the selected variables in a more manageable "full model" regression analysis, and appropriate "reduced models" developed to test particular hypotheses. It is likely that "full models" with pertinent interaction terms will procure the kinds of parameters needed in dynamic simulation. There would seem to be no a priori reason why economic and noneconomic data could not be included in regression analysis, thus providing additional clues to formulating and interrelating these two types of information in simulation modeling. There remains, however, a need to clarify the contrast between weighting economic data by regression and doing so by the Koopmans structural equation in which the estimated parameters are not regression weights, but are "elasticity coefficients". (See van de Greer, "Introduction to Multivariate Analysis for the Social Sciences", W. H. Freeman and Company, 1971.) It may be that including economic data in regression analysis will be useful to help identify the important variables, whose levels need to be included in simulation models, but to use elasticity coefficients, rather than regression coefficients in rate equations. The two sets of weights are contrasted by the different optimizing functions defining their estimation. (To add to the confusion, both sets of weights are called b-weights.)

Much has been indicated in the earlier discussions of the MISOE capabilities, subsystems, and analysis controls for inferential error bearing on static analysis across subsystems. Such analysis presumes the interconnectability of data across subsystems, ensured by logistics of data collection, tagging, and cross-referenceability in the information storage and retrieval system. Need for interconnectability applies to noncompletors as well as completors, and is further required for dynamic simulation as well as in static analysis.

Analysis performed in support of overall agency management decisions will generally be rather gross and limited by the kinds and quality of information available from other agencies. Analysis performed in support of management decisions over all education will involve both gross and specific data. Comparative analysis in terms of data types common to academic, general, and occupational education (i.e., input, product, and output) will be required. Analyses performed in support of particular program management, for LEA's, and for general occupational education management will be more specific and detailed with input-process-product-impact forms. It is anticipated that these remarks will apply to both static and dynamic analyses.

As soon as possible with available interfaced data, full model regressions should be set up and completed, rather than waiting for specific inquiries, so that information from the regression analyses can be used to identify important parameters. This information will also aid the staff in its further development of MISOE, in its interaction with managers in formulating inquiries to the system, and in formulating simulation models.

This completes the discussion of noneconomic, descriptive analyses in static space. The next part turns to the topic of simulation, especially in dynamic space, but will also include certain alternative considerations for analysis.

Part Five. Simulation Models

Introduction

This part considers the extensive staff commitment to simulation as a major analytic aspect of MISOE. The commitment, as expressed in Occasional Paper No. 3 and 6, is focused on the Forrester type models for dynamic simulation. Such models do, indeed, have a serious claim to be useful for MISOE, but certain limitations and possibly unknown characteristics of such models and their attendant applications in the MISOE context suggest an excessive staff emphasis thereon. Moreover, some types of inquiries can be anticipated for which static analysis or linear programming would be indicated and adequate.

For MISOE to have the general and flexible capability envisioned for the system, serious staff consideration should be given to static simulation, to linear programming, and to other kinds of solutions for some kinds of problems, and to other possible capabilities discussed in the extensive literature of operations research and of econometrics. To be sure, much of this additional capability, including the necessary software, will be available for economic analysis, but the point is that these tools may be also used for noneconomic analysis and for analysis which combines the economic and noneconomic concerns. It is in problems with strong nonrecursive, nonlinear, and temporal flow features where dynamic simulation will be most clearly indicated.

It may be instructive to consider an example of an inquiry which does not require dynamic simulation for its solution. The manager of occupational education wants to increase the quality of automechanics without changing the numbers or kinds of students entering the automechanics program. He conceives his problem as one of increasing the number of students with a Guttman product mix of "5". How is he to do this? MISOE approaches the problem by doing a stepwise regression analysis using the Guttman product mix score as the dependent variable, controlling for input, and using process data as independent variables. The students in all schools in the sample having automechanics programs are the observation units.

For simplicity in a didactic example, suppose two process variables were shown to predict the product variable: X_1, the square feet of floor space in the automechanics laboratory attended by the student, with a moderately positive regression weight; and, X_2, the number of students on an engine, with a larger, but negative regression weight. At this point, we know the critical variables for the manager to manipulate and, more specifically, that he will get more students with high product scores if he provides more engines in the laboratories, or if he provides more floor space (presumably so students working on adjacent engines are not bumping into each other). It also appears that providing more engines will be more effective in the quality of the student product than providing more floor space.

While helpful, this is quite inadequate. The formulation of the original inquiry was vague with respect to the nature or extent of the increase in product "5" students. Nor was any cost constraint imposed. Note, too, that one of the variables that "made a difference" (floor space) constrains the other, i.e., you can provide additional engines if you have enough floor space for them. It is doubtful that the manager would have thought of the latter until the "important" variables had been identified by the regression analysis. Even if he had defined the problem as getting a specified increase in the number of product "5" students for least cost, MISOE would now know which economic variables were most relevant (e.g., engine costs, costs of adding wings to school buildings, etc.). With such a least cost formulation and clues to the relevant variables, MISOE might recognise this as a linear programming problem, to be solved with the simplex algorithm, minimizing the total cost under the constraint that the number of square feet of floor space for a given number of engines is more than some specified constant.

The solution (if it exists) would specify the optimal levels of X_1 and X_2 in the sense defined.*

If the manager's inquiry were purely exploratory about the change in product mix distribution, preparatory to asking the cost of a specified increased product, MISOE could provide distributions of predicted product scores under status quo and under manipulated changes in X_1 and X_2. The distributions would be grouped by cutting scores defined by equicentile conversion against the actual status quo distribution (to allow for the regression effect) or by converting status quo predicted scores into stanine form. In effect, this would be a kind of static simulation of the effects of manipulated changes in X_1 and X_2 on the product distribution.

Dynamic simulation would be required if the manager's inquiry were made at a more sophisticated level. For example, the manager might specify that certain changes were to be made and might want to know how long it would take to reach the output distribution sought; or he might have to weight his decisions about automechanics in a context of similar decisions in other programs, or in regard to other outcomes for automechanics. It is also quite conceivable that some "alumni" from the automechanics programs become supervisors of future students either as teachers in occupational education or as supervisors in work-study programs and that process information turns up "important". With such temporal, mutually constraining, or feedback complications as these, dynamic simulation might well be necessary.

Before passing to a more detailed consideration of dynamic simulation, the attention of the staff is called to a linear programming approach to "assigning personnel to jobs". The logic of formulation and analysis is quite general so that "counseling" or "classification" can be substituted for "assigning" and "training channels" or "programs" can be substituted for "jobs". Usually, some produc or productivity measure is maximized rather than costs minimized.* Some useful references to such analysis, potentially useful to MISOE in dealing with inquiries about matching student mixes to programs, are:

---

*In this case, it is rather obvious without such analysis that providing more engines is less costly than adding wings to school buildings; the manager would do so up to the present floor space limitations. Such would not be the case, generally, with more variables, or with costs positively related to "importance".

"Methods of Solving Some Personnel-Classification Problems", D. F. Votaw, Jr. Psychometrika, 17, No. 3, 1952.

"Assignment of Personnel to Jobs", D. F. Votaw, Jr. and John T. Dailey, Research Bulletin 52-24, Air Force Personnel Laboratory, Lackland AFB, Texas, August, 1952.

"An Approximation Method of Solving the Personnel Assignment Problem", D. F. Votaw, Jr., and John M. Leiman, Technical Memorandum 56-14, Air Force Personnel Laboratory, Lackland AFB, Texas, July, 1956.

In addition to these references, a mimeo paper, "The Counseling-Assignment Problem", by Joe H. Ward, Jr. at the Personnel Laboratory at Lackland, and a Master's thesis by Donald Fink, presumably available from the Engineering Science School or library at Johns Hopkins University, are relevant. Most of this literature was developed for the Air Force and its personnel and training problems; there, matching personnel with programs involves quotas to be filled and constraints on the number of training slots available.

General Consideration of Dynamic Simulation

The next few sections will discuss various issues concerning dynamic simulation using Forrester type of formulation and the Dynamo capability. In this section we consider some general features of dynamic simulation with emphasis on kinds of models and on the flowcharting formulation of models. In subsequent sections, we consider equations and data sources, inferential errors, and other issues that have arisen in staff discussion.

Dynamic simulation models may be either general (gross) or specific (fine).

---

*The objective function could be to maximize the productivity over cost rati

This distinction may refer either to the order of the model (number of levels or rates) or to the magnitude of the time units. Inquiries from a state level manager over agencies or over all education are likely to involve relatively gross models, at least initially. More elaborate, fine-structured models may be involved, however, at those levels as managers become more aware of the importance of details in subsystems. Somewhat finer simulation models may be anticipated to answer inquiries from regional managers over educational sectors and programs. Model complexity is obviously a function of the number of conservative subsystems included, such as those dealing with programs, personnel, or costs.

Discussions with the staff revealed an expectation that some dynamic simulation models could be predesigned for call with specified parameters. In so far as this is practical, i.e., certain completely general and specificable models can be developed for clearly anticipated general forms of inquiries, the notion is an attractive one. It would seem more likely that variations in the detailed nature of the inquiries received by MISOE will imply variations in the details of the simulation flow charts and equations. If this is correct, much greater flexibility will be needed in formulating dynamic simulation models than one would have from a small set of prepackaged models. The latter in flow chart form may be initially helpful as a communication device with managers, and as a nucleus chart for the staff to elaborate in formulating specific models for specific inquiries.

As the staff accumulates experience in designing models for answering specific inquiries, portions of these models may be used as modules, which can be put together in various ways to form the initial flow charts for future inquiries. In this approach the same level and rate equations may be used in the new models wherever those levels and rates are not changed by their connections to other levels and rates.* Some types of modules that may thus develop over time, and may be repeatedly used include student flow subsystems, economic allocation subsystems, and certain kinds of information loops. It is likely that different modules developed in connection with an inquiry from some level of management will be most useful for inquiries coming from the same management level. How feasible or how helpful such a modular approach may be in MISOE requires further consideration and, possibly, actual operating experience. It is suggested here as a compromise between having a repertoire of a few general models and having to derive ad hoc models from scratch for every inquiry.

---

*This suggests that the user-defined MACROS in DYNAMO may be useful to MISOE.

Simulation runs with a given model may be classified as runs of the status quo, or as runs involving promulgated changes. Status quo runs should be made with any general model which may be feasible; it is likely that they will be the first runs with *any* model, to establish the behavior of the system as a base for comparing the results of any changes. Most changes may be expected to imply changes in parameter cards initiating levels and constants.without requiring any changes in the flow charts. It is conceivable, however, that a sophisticated manager may promulgate changes which will change the flow charts (e.g., he may decide he wants to use additional available information to influence one of the rates). Thus, the interaction between a manager and MISOE may stimulate his thinking *after* he has seen the results of status quo simulation, or of runs reflecting simpler changes.

Equations and Data Sources

This section considers the levels, rates, and rate modifiers in dynamic simulation. For each we consider how they enter a flow chart, how corresponding equations are formulated, and what kinds of data are to be retrieved from the information system. Some potential uses of DYNAMO functions will also be considered.

The *level* variables in a flow chart will come first of all from the concerns expressed in an inquiry. Where these involve different kinds of data (in terms of observation units, or economic vs. noneconomic data types), the flow chart should reflect these as nucleus levels for different subsystems, connectable *within* subsystems with level variables of the same kind, and connectable *between* subsystems with information links. The nucleus levels within a subsystem should then be supplemented with level variables that are inputs and outputs for each nucleus level. For every level variable, all of what flows in and out of that level should appear either as other level variables in that subsystem, or as sources and sinks. Moreover, if a level variable consists of several categories (e.g., the number of good citizens includes those entering, already in, and leaving an educational channel), and if it is necessary to keep track of such categories, one should ensure that all categories are accounted for in the definitions of levels with appropriate inputs and outputs, and that categories selected for explicit attention be mutually exclusive and without direct flow between them. Violations of these principles may lead to awkward or even inaccurate flow charts.

Values for the level variables will typically be frequency counts, or averages over some defined aggregation, ratios, or probabilities. Ratios are often called "rates" whether or not they are with respect to time. Those not with respect to time must be either levels or constants in dynamic simulation; those with respect to time may be levels, constants, or rates in the sense the latter term is used in dynamic simulation. Probabilities are ratios, expressing relative frequencies. Initial values of level variables for a simulation run are set by type N equations punched into DYNAMO control card. There must be a level equation for every level in the system (except for sources and sinks). Each equation will be of the form:

L Level.K = Level.J + DT (Sum of all RATES.JK controlling flow into the level minus the sum of all RATES.JK controlling flow out of the level)

There must be as many rate terms in the parenthesis as there are input and output channels to and from the level. DT is the simulation time unit, not the units of time in which rates are measured. Such units (including that of DT) must be consistent, conversion constants being used to ensure the consistency. All sources and sinks are indefinitely large and unspecified levels, but should appear in the flow diagram, connected to definite, specified levels by rate symbols and rate equations.

A rate symbol must appear between any two level variables in a flow diagram; a given level variable will have as many rate symbols attached as there are levels in direct connection with the given level. The rate *values* are not given directly to the computer as such, but are supplied through the rate equations and their modifiers; these, in turn, *may* be constants stating the rates directly, but usually are not, because the rates will not, in general, be constant throughout a simulation run but modified by the dynamics of the system. All factors, levels, constants, or other rates which can affect a given rate must be connected to the symbol for that rate by information lines in the flow diagram. Failure to do so may lead to confusion and to incomplete rate equations.

The formulation of rate equations poses the greatest challenge to the analyst; he cannot rely on the source of an inquiry to provide the factors that might affect rates, but must imagine, or determine in static analysis what is important and ensure its representation in the flow diagram. Morover, he must ensure that he has appropriate information about how factors affect rates. The curve fitting of trends data or introduction of tabulated functions may supply some of the necessary data expressive of such relations. More likely, though, the analyst will have to write a tentative, gross rate equation and then write auxiliary modifying equations that elaborate the major terms in the rate equation. When these auxiliary functions are evaluated and substituted back into the rate equations by DYNAMO, the rate equations are then fully specified and solved.

The rate equations are what give the simulation its dynamic aspect and all rate equations must have the basic common time unit of interest in the denominator. The general form of a rate equation is:

R Rate.KL = a function of (levels K and constants)/time

The levels and constants may be information about levels and constants from other subsystems having different observation and flow units, or there may be many of them impinging on a given rate, so that auxiliary modifying equations must be used. Modifying equations may take any form, provided that, when substituted back into the rate equation, they jointly preserve the unit dimensions as well as algebraic form. This principle not only checks the consistency of the equations, but also guides formulation of a complete and dimensionally consistent set of auxiliary equations.

The constants appearing in rate and auxiliary equations are specified as constant equations of the form C=k, where k is some value supplied from static space observation, analysis, or computation in static space. For example, they may be regression weights, partitioned variances, or ratios between retrieved aggregates. They may also be conversion constants, either in the sense of converting units of the same kind (a metric conversion) or in the sense of converting flow units in one subsystem to those in another (e.g., $/person).

Constants may also represent delay or adjustment times required for some information to feedback to affect a rate. It is conceivable that such times may be variable, rather than constant, and depend on system dynamics. Where this is the case, the delay time would be formulated as a level variable connected by rates to other levels affecting it, with the appropriate information loops indicated in the flow diagram. The delay functions provided in DYNAMO should be useful for the more common exponential delays, DLINF1 and DLINF3 for information delays, and DELAY3 for conservative flows of personnel and resources.

Goals and constraints may appear in rate equations, either as constants or variables, but are more likely to appear in auxiliary modifying equations. If variable, they would have to be treated in a manner similar to that for variable delays, as indicated above. Generally, the difference between a goal, or constraint, and the actual level of a variable would appear as a factor in a rate or modifying equation. To maintain dimensional consistency, it may be necessary to express such a difference as a ratio of the difference to the goal (or constraint).

This would be the case where such a function were to be multiplied in the rate equation by a factor already having the proper dimensions for the rate function; the factor for the difference between the actual and goal level should then be dimensionless to preserve those dimensions.

Although somewhat speculative at this point, it may be instructive to envision some potential uses of DYNAMO functions described in Chapter 8 of Forrester's Principles of Systems. The computational functions: SQRT, EXP, and LOGN, might be required if some equation involving these functions were fitted to data in static space; for noneconomic data, at least, this does not seem to be likely. The interpolation functions might be needed to obtain a constant for an auxiliary equation from a table of constants dependent on the value of some level, and the value sought is not tabulated.

The STEP function might be quite useful if some subsystem, not now connected, were anticipated to come into play at some specified future date. E. g., one wants to show a legislator what happens if requested funds become available next year instead of five years from now, the result may be different and the effect lag may not differ by 4 years. It might also be useful in a situation like that described on page 60, where the manager decides to buy more engines until present floor space constrains him; he may decide that the product quality is still not good enough and then decide to manipulate both floor space and engines.

The RAMP function might come in where the legislature decided to start funding a new program at a certain level and indicated that it would probably increase the funding steadily for several years. It is not clear how the trigonometric functions might be useful.

The noise generators should be useful in studying the effects of random variations in system parameters on a simulation model. Where mean values are used to initiate levels or as constants which enter rate equations, their standard deviations, standard errors of measurement, or sampling errors, with the mean and

NORMRN function modifying rates could be used to introduce these variations into dynamic simulation. Such variations may be considered as part of the "reality" being simulated, or may be used to estimate error effects on simulation runs.

The logical functions should prove quite useful, especially where rates will be functions of comparative magnitudes not otherwise expressed as ratios or differences. For example, the difference between a level and some goal may be a term in a rate equation; when the level surpasses the goal, it may be desirable not to let the negative difference affect the rate, but for the difference term to be evaluated as zero. This can be accomplished by MAX (or MIN, reversing P and Q parameters), by making the difference term equal to Q: MAX (0, G-L). CLIP will be more frequently useful, especially for allocation and resource limitation controls on rates. For example, the rate at which engines can be added to the auto-mechanic laboratories depend on dollars available for engines, but only as long as the amount available exceeds the cost of a single engine: R=f (CLIP DAV.K, 0, DAV.K, DPE), where DAV.K is dollars available for engines, and DPE is a constant cost per engine. SWITCH is a specialized form of CLIP.

Inferential Errors in Dynamic Simulation

This section considers various sources of error in dynamic simulation. Presumably, errors in initiating values may result in a distorted picture of the system after a period of simulated change. Unless the operations research literature or studies by persons working extensively with dynamic simulation exist, showing the effects of error, MISOE should conduct sensitivity studies to clarify this matter. It is plausable to expect the effects of error to be most severe at the beginning of a simulation run, and gradually diminish as the length of the run increases. Not only may this *not* be the case with some models, but we need to know how long it takes for error effects to become negligible, when it is the case. The answers to these questions may well be different for different kinds of models, in terms of their complexity, data types, order, number of information loops, delays, and actual rates. Presumably, the presence of negative feedback loops would be favorable to fast dampening of error effects, while positive loops may exacerbate them. In any case, in view of the strong commitment to use dynamic simulation in MISOE, it is strongly recommended that the staff search the technical literature and perform whatever necessary experiments are required to clarify these issues.

It is also necessary to develop a strong sensitivity to the prevention and recognition of modeling errors. For example, failure to consider all of the factors that might affect a rate or the use of an improper function will, in effect,model something other than one's hypothesis about reality. The diagnostic messages in DYNAMO, like those of most good compilers, will catch most errors that are violations of the DYNAMO language, whether these resulted from erroneous formulation of equations or keypunch errors. They will also indicate certain inconsistencies, lack of definition, and mathematical impossibilities. They will not, of course, identify errors of conceptualization, or tell you when a feedback loop should have been included to have a variable effect on a rate, instead of merely supplying a constant.

In typical examples of dynamic simulation the variables are expressed in clearly defined and well understood metrics (e.g., numbers of people or dollars, physical units of length, weight or time, electrical units of voltage, or capacity, or ratios of such units). Such units are readily convertable to other units of the same kind (feet to miles, months to years, pounds to tons, etc.) by linear conversion constants. It is not clear, however, whether the arbitrary metrics of psychological test scales and their various transformations, both linear and nonlinear can be used in the same way in dynamic similation. In any case, some kind of metric consistency with respect to raw scores vs. standard scores and the proper choice of regression weights (b or beta) must be insured when such data are used in dynamic simulation.

It is likely that the use of standard scores which can go negative (Z-scores) may give some trouble in the behavior of a system equation. This may be avoided by the use of T-scores formed by adding a constant to the Z-scores with or without their normalization (McCall's T). Where standard scores are used involving more than one group, the scores should be on a common normative base. One can convert simulation outcomes, if necessary, to within-group defined metrics, for communication of results to different program managers in case they are using program norms rather than statewide norms. The nature and extent of some of these problems requires further discussion, and some can be headed off by appropriate management of testing, and of the reporting of test results. Note that many commercial tests are normed on arbitrary customer samples, and may or may not be relevent metrics for MISOE purposes.

Initiating values of levels and constants, many of which come from sample data and analysis in static space, are subject to sampling errors and errors of measurement. A fuller discussion of sampling errors and their effect on computered statistics will be included in Occasional Paper No. 12. It is apparent that the use of a mean value for a level or a constant is associated with a standard deviation, so that there is some noise in these values when they are used in rate equations. Consideration should be given to the comparison of two simulation models, one of which ignores such variation, and the other of which takes it into account by means of the NORMRN function in DYNAMO. Similar considerations apply to the effects of measurement error in levels and constants. The correction for attenuation in correlational analysis in static space corrects certain relationships among variables, but not their observed values.

The total variance about an observed value may consist of "true" variation of some kind, sampling error, or measurement error. Attempts to minimize measurement error by setting high reliability requirements on observed variables, and to minimize sampling error by means to be considered in Occasional Paper No. 12, will reduce the seriousness of error effects in dynamic simulation. Nevertheless, we cannot be sure that they will be reduced below some unknown tolerance level in dynamic simulation. This is why we need more information about what that tolerance level might be for different kinds of simulation models.

Nonrandom bias in levels and constants, when unknown, may have more serious effects in dynamic simulation than the more nearly random variations discussed above. The control of bias in sampling will be discussed in Occasional Paper No. 12, through consideration of sampling logistics and weighting procedures. The present concern is with those simulation parameters (e.g., regression weights) resulting from correlational analysis, especially as a result of incomplete prediction. It may be that regression analysis will be much more useful for identifying important relationships than for supplying regression parameters to simulation. Actual values of dependent and independent variables will be available and normally should be used in simulation in preference to predicted values. One can use regression weights or partitioned variances where needed with greater confidence when the multiple correlation is high. Again, it may be worthwhile to conduct an experiment with a simulation model using data involving a full model regression where R^2 is high, and to compare the results with that of a similar model, ignoring some of the regression variables and using the weights recomputed on the corresponding reduced regression model.*

Because MISOE is pioneering the use of dynamic simulation with a mixture of physical, economic,and psychological data, it will probably not find the answers to these questions in the available literature. It must, therefore, be prepared to be a pioneer in facing up to some of the methodological issues which are presumably new in MISOE design and operation. It is not intended that these issues sidetract the main operational thrust envisioned for MISOE, but only to insure a high level of accuracy in the results of analysis fed back to management, and on which their policy decisions may be based.

---

*The weights for the dropped variables are zero.

A Pseudodynamic Model as Nonlinear Programming

This section responds to a particular staff request. The request consisted of ascertaining whether the following problem could be formulated and solved in dynamic simulation:

> The manager within a certain program (e.g., automechanics) wants to choose the least cost process mix which will transform a certain input mix into a certain product mix.

The staff particularly wanted an example at this management level (input-process-product). The following assumptions and attitudes were imposed on the attempted solution:

1. the example would be designed to raise issues for further discussion about the feasibility and techniques for handling such an inquiry via dynamic simulation methods;

2. the possibility of a static space or linear programming solution to such an inquiry would be ignored for the present;

3. the example would be set up as a comparison of two process mixes, submixes, or elements (whether human, physical, or organizational), in such a way that the model could be easily generalized to the comparison of more processes, or to additional input and output mixes;

4. the fundamental flow would be from an input level (IL) of the number of students with a certain input mix to the product level (PRODL) of the number of students with a specified product mix, with process and cost information moderating the rate equation corresponding to the flow of students through the process;

5. there would be a flow channel for each process and the rate equations would be formulated in such a way that there would be a null rate for all but the least cost channel;

6. the example would be kept as simple as possible without feedback

loops and explicit economic or other information subsystems; all information, however, must be available in storage or from static space analysis.

The "flow diagram" so formulated appears in Figure 1. The following rates are ignored:

R_{n+1} for the EHI source of input students,

R_{n+2} for those students with the "certain input mix" who go to some other output mix (OPML) by any process,

R_{n+3} for those with other output mixes going to the societal impact sink,

R_{n+4} for those with the specified product mix going to the societal impact sink, where n = the number of processes compared.

Note that the processes, themselves, do not appear explicitly in this diagram, except as labels on the flow routes, although information about process-product relations does. Note, also, that the two "processes" are not specified, but could be the use of laboratory vs. putting students through a cooperative work-study plan; or having older, more experienced, and higher salaried instructors vs. having younger, possibly more flexible, and lower paid instructors.

The "rate" associated with each process consists in part of an auxiliary function of two factors: the probability that some portion of IL will move to PRODL in process time and that it cost so much per student to do so. No regression information is required; only the student mix by product mix I/O table and the corresponding cost per student table. This auxiliary function expresses the basic flow rate in terms of the probable benefit and its cost, as an inverse cost-benefit ratio, all over process time. Note that this permits the two (or more) processes to require different lengths of time. The rate equation for each process is conceived as consisting of the auxiliary function, or basic rate, modified (i.e., multiplied) by a CLIP function, or as many such CLIP functions as there are other processes in the model. The purpose of the CLIP function, here, is to leave the

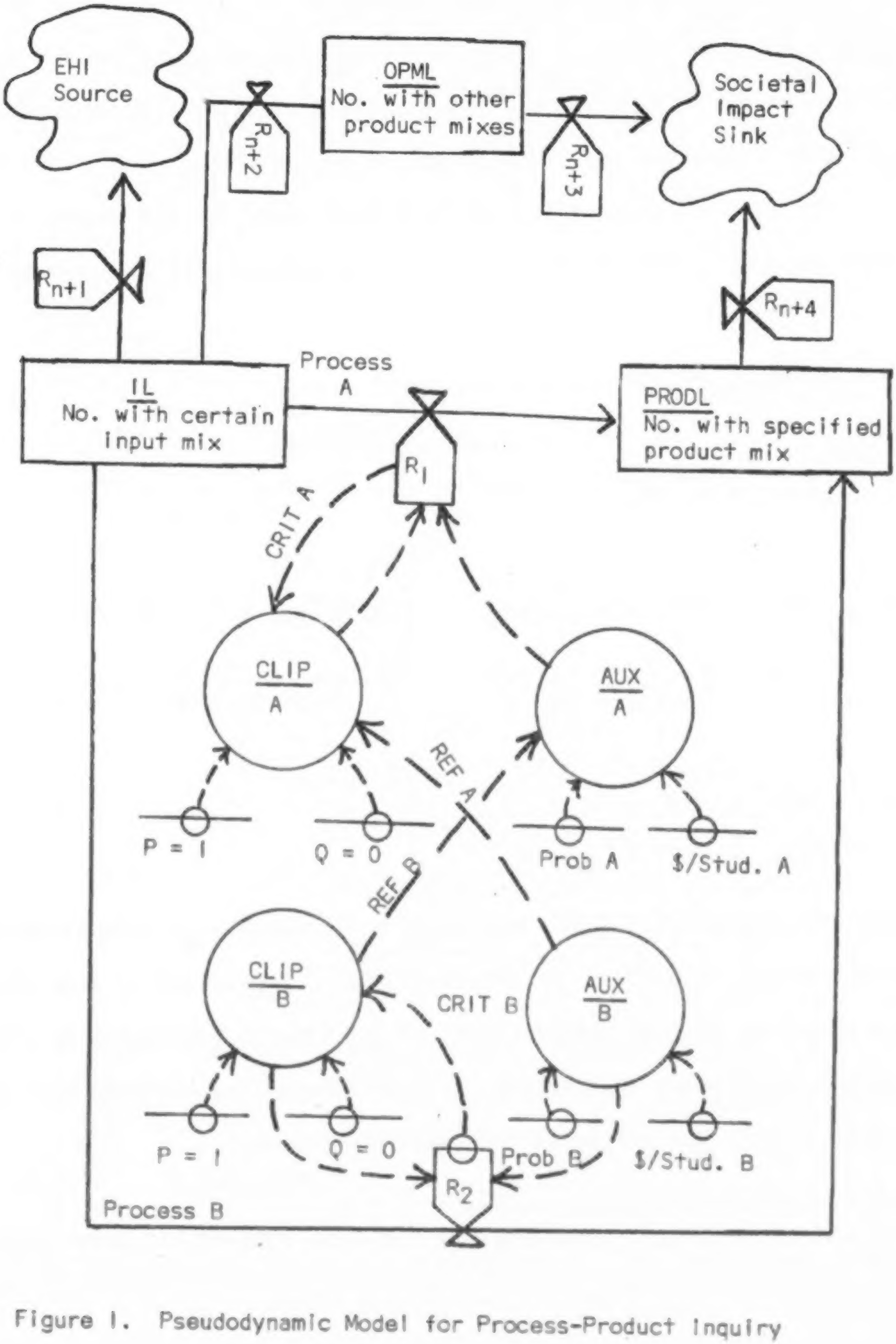
EHI Source
OPML
No. with other product mixes
Societal Impact Sink
R_{n+2}
R_{n+3}
R_{n+1}
R_{n+4}
Process A
IL
No. with certain input mix
PRODL
No. with specified product mix
R_1
CRIT A
CLIP
A
AUX
A
REF A
P = 1
Q = 0
Prob A
$/Stud. A
REF B
CLIP
B
CRIT B
AUX
B
P = 1
Q = 0
Prob B
$/Stud. B
R_2
Process B

Figure I. Pseudodynamic Model for Process-Product Inquiry

rate as computed by the auxiliary equation, multiplied by 1, if that process is already the one with the "best" cost-benefit ratio (it is only least-cost if all channels produce the same number of students with the given output mix for a given input number). Otherwise, CLIP changes the auxiliary function to a rate of zero. It is assumed that DYNAMO can solve the equations so formulated with a DT about 1/3 the length of the shortest process and can be made to print out the values of the rate functions at 3-6DT. The answer is to take the process with a "non-zero" rate.

The auxiliary and rate equations are:

$$AUX_A = \frac{\text{Prob}\ \ (\text{PROD/IL})^{*} \times \text{Studs}_{A/\$}}{\text{Process}_A\ \text{Time}} = \frac{\left(\frac{\text{IL-PRODL}}{\text{IL}}\right) \cdot \frac{\text{Studs}_A}{\$}}{\text{Process}_A\ \text{Time}}$$

$$AUX_B = \frac{\text{Prob}\ \ (\text{PROD/IL}^{*} \times \text{Studs}_{B/\$}}{\text{Process}_B\ \text{Time}} \qquad \frac{\left(\frac{\text{IL-PRODL}}{\text{IL}}\right) \cdot \frac{\text{Studs}_B}{}}{\text{Process}_B\ \text{Time}}$$

$$R_I = AUX_A \times \text{CLIP A}$$

$$R_2 = AUX_B \times \text{CLIP B}$$

Whatever else can be said about this example, it appears to have answered the manager's inquiry, as stated, but the formulation, regardless of flow diagram symbols, and equations of the "proper" form, is not dynamic simulation in the Forrester sense. It appears that dynamic simulation modeling concepts have been used to perform a kind of brute force nonlinear programming.

---

[*] Probability of a student obtaining the specified product mix given that he had the certain input mix.

It should be noted that this formulation does not define the least cost process mix, but only chooses the least expensive one, among those defined; presumably this could be done by direct comparison of total product costs. The particular class of inquiries involving process-product analyses, with relatively fixed program lengths, and constrained by fixed input and output would not seem to require dynamic simulation, unless embedded in the larger system implied by other student and product mixes, with the larger system including data over a longer time (e.g., impact data), and feedback loops. Moreover, there is a kind of cohort batch effect in the flow of students through a program, which can probably be ignored when program length is small compared to the process time for continous flows in a larger simulation model.

From the viewpoint of dynamic simulation, the present model is a degenerate one and the forcing of the rate values to be either positive or zero a gimmick. The manager's inquiry is perfectly reasonable, but should be soluble in static space by other procedures or models. The next two sections discuss two such possibilities.

A More Rational Approach

Another approach to the problem discussed in the last section is to formulate it in terms of vector differention, following the method described by Van de Geer (Introduction to Multivariate Analysis for the Social Sciences, W.H. Freeman and Company: San Francisco, 1971, p. 58-59).

1. Perform the regression analysis of "the certain product mix" on the process variables, using the students with certain input mixes as the unit of analysis.

   This gives the equation:

   > Predicted product level (PRODL) = $b'X+C$, which is converted to the form, $g = b'X+C - \text{PRODL} = b'X+K$. X is the process mix column vector sought; b is the column vector of regression

weights, C the regression constant (K = C-PRODL).

2. Assume that for each x_i process variable selected by the regression analysis the unit cost data are available (e.g., \$/sq. ft., \$/teacher contact time with student, etc.). The total product cost is given as a vector, Y, of the actual process variable costs:

   Y = VX, where V is a diagonal matrix of unit costs and X is the column vector of process variables, as above. (The grand total product cost is 1Y, where 1 is the unit row vector).

3. The cost function in vector form is linear. However, in order to apply the suggested procedure, we need it in bilinear form for minimizing under the constraint of relating process to product, as expressed in the regression solution. To obtain the bilinear form required, define the scalar $Y^* = Y'Y = X'V'VX$. Minimizing Y^* minimizes the sum of the squares of the actual costs of the process variables. Letting $V^* = V'V$, $Y^* = X'V^*X$, the bilinear form required for solution.

4. Set the function g (in step one above) to zero.

5. Define the auxiliary function:

   $F = Y^* - Ug = X'V^*X - \mu\,(b'X+K)$, where μ is a Lagrangian multiplier. Take partial derivatives of F with respect to the x_i, evaluate the Lagrangian multiplier, and solve for those values of process variables, that minimize the sum of the squares of the costs of those process variables most predictive of the stated product mix. It will be shown below that it is a minimum.

The above steps require a little further discussion. In step 1, the proper treatment of the control of the regression for "the certain input mix" is not entirely clear. If the input mix is a constant vector, input is already controlled; if the input mix is a class of input vectors, a possible treatment is to residualize the PRODL against it, and then predict the residualized PRODL from process variables. In step 2 there is the assumption that cost data are available for each process variable selected and that it is in, or transformable into, the proper form. The substitution, in step 3, of the sum of squares of actual process costs for the total product cost as the function to be minimized seems plausible, if somewhat forced. If a linear function is used, the desired x_i variables differentiate out and a trivial solution results. Using the squares of process variable costs will tend to depress the x_i values for the most expensive process variables, which has some intuitive appeal.

The setting of the g function to zero in step 4 amounts to using the actual, rather than the predicted value of PRODL. If the residualizing against input is used, it amounts to using the actual residualized PRODL rather than the predicted residual. The validity of doing this depends on the actual value of R in the regression solution.

It remains to expand the solution in step 5 more explicitly, and to show that the solution is a minimum:

$\sigma F/\sigma X = 2V^*X - Ub = 0$

$X = \mu V^{*-1}b/2$, but μ is as yet unknown. Since $b'X = -K$ when $g = 0$, $b'X$ is known, and can be set $= \mu\, b'V^{-1}b/2$. Solving for μ, $\mu = 2\ (b'V^{*-1}b)$. Substituting μ back in the equation for X, $X = V^{*-1}b\ (b'V^{*-1})^{-1}$, the column vector of values of the process variables that minimize Y^*.

To show that the solution is a minimum, evaluate the second derivative of the F function; it is $2V^*$. Since the squares of process variable costs are positive, the second derivative is positive, and therefore, the solution is a minimum.

The feasibility and generality of this solution for MISOE should be discussed further and compared with other possible alternatives. It appears to be a form of nonlinear programming.

A Linear Programming Solution

The approach in the previous section started with the recognition that the vector differentiation of a bilinear function could be useful in such a problem, but in step 3, the cost function was redefined to ensure a match to the model. The essentially linear nature of the problem, both in the regression function and in the original cost function, suggests the possible use of linear programming as a solution. In order to formulate a problem by linear programming, it is necessary that the basic concepts and assumptions in such an approach are met. (Dantzig, George B., Linear Programming and Extensions. Princeton, N.J.: Princeton University Press, 1963.) With little imagination, the process variables may be thought of as "black box" activities, and the students as items flowing from input to product spaces. Imagination becomes somewhat strained, however, with respect to the concept that activity levels (i.e., values of process variables) are changed by flows into and out of the "activities," and by the "proportionality" assumption that a doubled flow (of students) doubles the activity levels. There is also some strain with respect to the additivity assumption, as usually interpreted in linear programming problems. Nevertheless, the assumption that process levels are nonnegative is readily met (with linear transformations on the variables, if necessary), and the expression of conservation of a precious item (money, in this case) in a linear objective function is applicable. Moreover, it is possible to formulate a set of equations in our problem, which have the same mathematical form as those in a typical linear programming problem. These equations are:

1. $Y_{res} = BX$, where Y_{res} is a column vector of PRODL, residualized

against input; B is a diagonal matrix of regression weights; and, X is the column vector of process variable levels. The B matrix may be replaced with a matrix, C, of variance contributions, which may not be a diagonal matrix. In either case this subsystem of equations corresponds in form to the material balance equations of linear programming.

2. z(min) = vX, where z is the product cost to be minimized, v is the row vector of unit process costs, and X is the column vector of process levels, as above. This is the objective function.
3. $x_i \geq 0$ expresses the nonnegativity restriction, which must apply to both the regression and current models in this formulation.

An important question is whether, in a given application, this system of equations can be solved with the simplex algorithm. Moreover, the concerns expressed throughout this paper about model validity and inferential error are also applicable here.

With this, we now have three formulations of the original inquiry:

1. pseudodynamic, which told which of two process mixes already defined was less expensive, but which did not define the mix in terms of levels;
2. nonlinear (i.e., bilinear), which yielded a process mix minimizing the sum of squared costs rather than the total product cost, and
3. the present linear approach, which defines a process mix minimizing the total product cost, as asked, if the simplex solution exists.

For the kind of problem raised here, it is recommended that the linear approach be tried first, and if it fails, that the nonlinear approach be tried. It is probably best to write off the pseudodynamic approach as a learning experience. Again the feasibility and generality of the linear approach for MISOE problems should be discussed further. It is likely that minor variations in the way the original inquiry is formulated, while still focused on the goal of finding the least cost process mix, will render a better match with linear programming concepts and assumptions. By analogy with the transportation type of problem, linear programming would appear to be feasible for a class of inquiries characterized by different student input mixes going through different process channels to different product mixes. The different student mixes are the items, the different and unknown numbers of students of each type going through each process are the activities. Here, the processes, per se, are "black boxes," available inputs and output quotas are constraints, and the objective function is to minimize the overall system cost. The relevant process data and associated process costs presumably provide the coefficients in the system of equations. That is why the personnel assignment problem in the military, referred to in an earlier section, was amenable to the linear programming approach.

Part Six. Epilogue

This short, final part consists of a few general, summarizing statements, or "conclusions" as follows:

1. MISOE needs a highly varied repertoire of general models and algorithms.

2. Regression is a powerful tool. It may solve some problems directly, or with little further effort in static space. It may be used to generate expectancy tables of the type described by the author for counseling and/or admissions problems (Creager, J.A. "Use of Research Results in Matching Students and Colleges," The Journal of College Student Personnel, Sept. 1968). Regression is most likely to be useful in identifying the important variables for other analyses and to give some information about the "relative importance" of those selected. Regression parameters may also be useful as simulation parameters, but this appears to be less likely than originally thought.

3. MISOE needs to maintain flexibility of options until more is known about the relative frequency of inquiry types from various levels of management.

4. There may be some limitations on the utility of dynamic simulation, because the conditions for its use are not yet completely specified, and its sensitivities to various kinds of error are not yet adequately documented. This whole area needs further study.

5. MISOE needs to engage in some "shakedown" experiences before becoming fully operational, perhaps including a pilot, partial implementation period. Both ethical and pragmatic considerations require great attention in MISOE development and implementation to the sources and control of inferential errors in the application of all analysis models.

6. MISOE needs to assess more clearly the utility of linear programming and other models for integrating the economic and noneconomic aspects of analysis.

7. The current status of MISOF represents a vision of great potential use as a system in support of management. Many problems have been faced and worked out, either in whole or in part. Much remains yet to be resolved before initial implementation; some matters will be resolved in the context of operating experience.

APPENDIX A

TECHNICAL REPORTS FOR MISOE

3-102

HEADQUARTERS
6560TH RESEARCH AND DEVELOPMENT GROUP
(PERSONNEL RESEARCH LABORATORY)
HUMAN RESOURCES RESEARCH CENTER
LACKLAND AIR FORCE BASE
San Antonio, Texas

STAFF RESEARCH MEMORANDUM
Project: 503-001-0016

2 September 1953

STUDIES IN METHODOLOGY

II. EFFICACY OF THE UNIVARIATE FORMULAS FOR CORRECTING FOR RESTRICTION OF RANGE

John A. Creager

One of the frequently encountered problems in the treatment and interpretation of psychological data is that of correcting a correlation coefficient for restriction of range. This memorandum is concerned with some characteristics of the correction formulas as they are applied to Pearson coefficients, where both variables are continuous and normally distributed in the unselected population. For the case of univariate selection, three basic formulas are available (AAF Research Report #3, "Research Problems and Techniques", pp. 63 - 68; Cf. Thorndike, R. L., Personnel Selection, pp. 169 - 176):

Correction Formula I: $$R_{12} = \sqrt{1 - \frac{s_2^2}{S_2^2}(1 - r_{12})} \quad ,$$

where R_{12} is the corrected correlation coefficient, r_{12} is the available correlation in a sample restricted on variable 1, s_2 is the standard deviation of the indirectly restricted variable 2 in the restricted group, and S_2 is the standard deviation of the indirectly restricted variable 2 in the unrestricted group.

Correction Formula II: $$R_{12} = \frac{r_{12}\frac{S_1}{s_1}}{\sqrt{1 - r_{12}^2 + r_{12}^2\frac{S_1^2}{s_1^2}}} \quad ,$$

where the symbols have the same meaning as in Correction Formula I, except that the standard deviations are available for the directly restricted variable 1.

Correction Formula III: $$R_{12} = \frac{r_{12} + r_{13}\, r_{23}\left(\frac{S_3^2}{s_3^2} - 1\right)}{\sqrt{1 + r_{13}^2\left(\frac{S_3^2}{s_3^2} - 1\right)}\sqrt{1 + r_{23}^2\left(\frac{S_3^2}{s_3^2} - 1\right)}}$$

---

[*] HRRC Staff Research Memoranda are informal papers intended to record opinions and preliminary reports of studies. They may be expanded, modified, or withdrawn at any time and hence are not suitable for inclusion or reference in more permanent reports of a scientific or technical character.

where direct restriction occurred on variable 3, and both variables 1 and 2 had been indirectly restricted by their correlations with the directly restricted variable 3.

For the case of simultaneous correction of many coefficients, as in a correlation matrix, with either univariate or multivariate selection, matrix formulations for correction of range restriction are available, and will be discussed in a subsequent memorandum.

The basic assumptions underlying these various fomulations for correction of range restriction are:

a. The regressions of indirectly selected variables upon directly selected variables are linear and homoscedastic in the unrestricted population.

b. The slopes of the regression lines are unaltered by selection.

c. The standard error of estimate of indirectly selected variables from directly selected variables is unaltered by selection.

d. The partial correlation between indirectly selected variables is unaltered by selection.

If the first two assumptions are met, the last two will generally be met also. In Gulliksen's development of these formulas, no explicit assumption is made regarding normality of the distributions of the two variables. Extreme deviations from normality will, however, make it quite difficult to meet the stated assumptions.

In practice, the application of these formulas assumes that direct selection has occurred entirely on known and measured variables. Thus technical school criterion data may have been subjected to sources of selection other than the explicit requirements for career guidance and assignment. In such a case the correction would generally err on the side of conservatism, i.e. the formulas would underestimate correlation for the unselected population.

The restriction imposed by the meeting of priorities in fulfilling quotas may also effect the applicability of these formulas. For example, 1000 men may have been assigned to two technical schools, A and B, on the basis of the same stanine cut-off. Suppose then that 600 men qualify with a stanine of five or greater. Suppose further that School A requires 250 men with top priority and School B must take what is left. The assignment of the 600 men from two schools, in terms of their stanine scores, would look as follows:

Stanine	A	B	
9	40	0	
8	70	0	
7	120	0	
6	20	150	
5	0	200	
Σ	250	350	(N = 600)

While this is a rather extreme example, it is obvious that the assumptions listed above will have been violated.

The present study was undertaken to test the validity of the univariate correction formulas under conditions meeting the assumptions, and to ascertain the effect, if any, of linear dependence between directly and indirectly restricted variables.

These studies were carried out using an experimental population, previously prepared using punched card procedures, and described in the first memorandum of this series (Studies in Methodology - I. Description of an Experimental Domain for Methodological Studies). The theoretical correlation and distribution statistics for the unrestricted population are given in Table 1. A random sample of 500 cases was obtained and two restricted samples prepared as follows:

> Restricted Sample I. The 300 cases having the highest score on the composite test, #11, were selected from the random sample of 500 cases. This provided a sample of 300 cases directly restricted only on a composite score, thus simulating the conditions resulting from selection on an aptitude index stanine of five or greater to obtain a pool of "qualified" men. The fulfillment of quotas may yield different results as previously discussed.
>
> Restricted Sample II. The 180 cases having the highest score on test #1 were selected from the random sample of 500 cases. This sample simulates a somewhat more severe restriction upon a single non-composite test.

The intercorrelations and distribution statistics for these two restricted samples are given in Tables 2 and 3, respectively.

With these restricted samples and the "true" population values available, certain questions concerning range restriction corrections can be answered. The first question was that of comparing Correction Formulas I and II for their efficacy in correcting a correlation between a directly restricted variable and an indirectly restricted variable. The difference between the formulas is dependent upon the available information, i.e. whether standard deviations are known for the indirectly restricted variable (Formula I) or for the directly restricted variable (Formula II). In many practical situations both standard deviations are known and, if both formulas are applied to correcting the same restricted coefficient, appreciable discrepancies may be occasionally noted in the corrected coefficients. Table 4 shows the errors (corrected coefficients minus "true" coefficients) obtained by applying Correction Formulas I and II to correlations involving the directly selected variable. Somewhat larger errors are encountered with Formula I than with Formula II. The greater errors for restricted Sample II are due to the smaller size of the sample, i.e. due to sampling errors in the restricted correlations. It is also apparent that Formula I is somewhat more sensitive to such errors. The highest errors occurred where the restricted correlations were negative. The practical conclusion is that, where standard deviations are known for the directly restricted variable, Correction Formula II should be used in preference to Formula I.

Attention is called to the fact that the corrections in restricted Sample I are valid for tests which are weighted into the restricted composites. Hence, Correction Formulas I and II do not appear to be invalidated by linear dependence between directly and indirectly restricted variables.

Another questions considered involves the efficacy of Formula III, for correcting intercorrelations among indirectly selected variables. This was tested only for restricted Sample I. Each correlation was corrected individually by Formula III and the resulting coefficients compared with the "true" values from Table I to yield the error matrix given in Table 5. This matrix has been augmented by a row vector consisting of the errors resulting from applying Correction Formula II to the correlations involving the directly restricted composite test #II. These errors are small enough to be attributed to sampling errors in the original restricted sample (treated as a random sample of 300 cases from a restricted population). It is also apparent that Correction Formula III is valid for those particular tests which are components of the explicitly restricted composite.

It should be emphasized that the present study dealt exclusively with univariate selection, involving Pearson coefficients of correlations, where both variables are continuous and normally distributed, and where practically all of the restriction occurred only on the single directly restricted variable.

TABLE 1

Theoretical intercorrelations* and Distribution Statistics for Experimental Domain (N ⟶ ∞)

Test	1	2	3	4	5	6	7	8	9	10	mean	S.D.
1											8.100	1.96
2	000										8.160	1.96
3	640	400									8.160	1.96
4	560	000	560								8.550	1.96
5	000	320	200	350							9.680	1.96
6	240	480	540	410	520						10.870	1.96
7	400	560	750	350	280	570					9.345	1.96
8	260	000	240	610	560	410	150				8.100	1.96
9	000	240	150	300	540	420	210	480			9.680	1.96
10	320	320	520	460	440	520	480	440	360		11.100	1.96
**11	444	346	720	693	741	670	580	763	495	639	85.900	14.483

*Decimal points have been omitted
**Test 11 is a raw score composite = 4 x (Test 3) + 3 x (Test 5) + 3 x (Test 8)

[illegible] 2

Intercorrelations* and Distribution Statistics for Restricted Sample #1 (N = 300)**

Test	1	2	3	4	5	6	7	8	9	10	mean	S.D.
1											8.744	1.897
2	-214										8.542	1.814
3	575	293									9.159	1.666
4	470	-184	405								9.430	1.712
5	-306	213	-239	028							10.518	1.548
6	006	346	319	198	294						11.596	1.673
7	330	460	673	224	-042	374					10.103	1.737
8	007	-155	-138	369	335	183	-149				8.910	1.560
9	-172	121	-084	096	396	293	004	359			10.186	1.783
10	152	240	420	239	216	412	397	157	178		11.988	1.777
***11	272	247	559	508	530	485	401	593	330	499	94.915	8.962

*Decimal points have been omitted

**Restriction has occurred on Test #11.

***Test 11 is a raw score composite = 4.0 (Test 3) + 3.0 (Test 5) + 3.0 (Test 8)

TABLE 3

Intercorrelations* and Distribution Statistics for Restricted Sample #2 (N = 180)**

Test	1	2	3	4	5	6	7	8	9	10	Mean	S.D.
1											8.621	1.704
2	-067										8.124	1.880
3	571	391									8.375	1.793
4	513	-013	522								8.750	1.804
5	-095	358	185	271							9.563	1.958
6	185	445	575	425	515						10.739	1.927
7	360	550	720	275	293	520					9.384	1.875
8	119	-010	162	505	490	416	056				7.945	1.818
9	-096	219	061	189	538	378	127	465			9.667	1.929
10	313	358	533	441	422	552	495	405	304		11.207	1.962
***11	316	367	690	611	745	712	543	717	464	643	86.022	13.238

*Decimal points have been omitted
**Restriction has occurred on Test #1
***Test #11 is a raw score composite = 4.0 (Test 3) + 3.0 (Test 5) + 3.0 (Test 8)

TABLE 4

Errors in Correcting for Range Restriction Using Correction Formulas I and II*

	Restricted Sample #1**		Restricted Sample #2***	
Test	Error by Formula I	Error by Formula II	Error by Formula I	Error by Formula II
1	-08	-03	—	—
2	10	04	29	-08
****3	-01	02	02	-02
4	-04	00	05	01
****5	00	-03	11	-11
6	-01	00	02	-04
7	00	00	05	-01
****8	01	00	15	-10
9	02	00	20	-11
10	-02	04	-01	03
11	—	—	06	-09

*Decimal points have been omitted
**Restriction on Composite Test #11 (N = 300)
***Restriction on Test #1 (N = 180)
****Test weighted into Composite Test #11

TABLE 5

Complete Error Matrix Resulting From Individual Application of Formulas II and III to Restricted Sample #1 (N = 300)*

Test	1	2	3	4	5	6	7	8	9	10
1										
2	-095									
3	-008	000								
4	-010	-014	031							
5	-057	014	-025	-029						
× 6	-063	042	-015	016	-021					
7	025	-035	000	063	-057	-060				
8	-038	062	019	-036	-002	030	018			
9	-023	-019	137	-016	-022	014	-032	016		
10	-023	034	073	-028	010	057	061	-012	-013	
**11	-028	035	017	-003	-030	-063	-002	003	-003	042

*Decimal points have been omitted
**Restriction has occurred on Test #11, a composite involving Tests 3, 5, and 8

Note: σ_{r_0} for N = 300 is 0.058

3-103

HEADQUARTERS
6560TH RESEARCH AND DEVELOPMENT GROUP
(PERSONNEL RESEARCH LABORATORY)
HUMAN RESOURCES RESEARCH CENTER
LACKLAND AIR FORCE BASE
San Antonio, Texas

STAFF RESEARCH MEMORANDUM*
Project: 503-001-0016

2 September 1953

STUDIES IN METHODOLOGY

III. A NOTE ON THE MATRIC FORMULATIONS FOR CORRECTING FOR RANGE RESTRICTION

John A. Creager

The purpose of this memorandum is to consider certain aspects of the matric formulations for correcting measures of covariation for restriction of range. Such matric formulations have been given by Thorndike (AAF Research Report #3 "Research Problems and Techniques", p. 67; Cf. Personnel Selection, p. 176) in terms of correlation, and by Gulliksen (Theory of Mental Tests, pp. 158-171) in terms of covariances. Two purposes are served by these formulas: To permit simultaneous correction of the whole correlation matrix, and to permit corrections for multivariate restriction. The first problem arises in multiple regression studies and factor analyses of criterion data. The second problem may arise where criterion data are under investigation and selection occurred simultaneously on an aptitude index and on a test being considered for addition to the classification battery. The latter situation arose in connection with studies of Radio Operator trainee selection.

The present inquiry is concerned with two major problems. The first involves a clarification and interpretation of the matric formulations with special attention to the relationships among various formulas. The second problem involves the empirical study of the efficacy of the matric formulas.

The assumptions underlying the matric formulas are the same as those for the three univariate selection formulas discussed in a previous memorandum (Studies in Methodology - II. Efficacy of the Univariate Formulas for Correcting for Restriction of Range). In addition the number of variables must be identical in the restricted and unrestricted groups; indeed, the variables themselves must be identical in both groups for the matric operations to have any meaning.

Prior to consideration of the first problem, it is necessary to clarify the notation systems used by Thorndike and Gulliksen, respectively, and to be sure that subscript notation is consistent for the subsequent discussion. Table 1 provides useful reference for this purpose.

---

*HRRC Staff Research Memoranda are informal papers intended to record opinions and preliminary reports of studies. They may be expanded, modified, or withdrawn at any time and hence are not suitable for inclusion or reference in more permanent reports of a scientific or technical character.

TABLE I

Notation Schemes for Thorndike and Gulliksen Matric Formulations for Correcting for Range Restriction*

Verbal Description	Thorndike Notation	Gulliksen Notation
1. Directly (Thorndike) restricted variables Explicitly (Gulliksen)	a	x
2. Indirectly (Thorndike) restricted variables Implicitly (Gulliksen)	x	y
3. Covariation Matrices	R, r (correlation matrices with unit diagonals)	C, c (covariance matrices with variance diagonals)
4. Diagonal Matrices of Standard Deviation Ratios, Σ/σ	H	-
5. Diagonal Matrices of Variance**	-	V, v
6. Matrices of Partial Regression Weights in the Restricted Groups***	b_{ax}	w_{xy}

*Capital letters are used for data in the unrestricted group. Lower case letters are used for data in the restricted group. Exception: H, a diagonal matrix, Σ/σ, involves data from both groups.

**These variance matrices were added by the writer to facilitate the translation of Gulliksen formulas to Thorndike formulas.

***Subscripts on all matrices refer to rows and columns, respectively, where a single subscript is used, a square matrix symmetrical across the diagonal may be assumed.

The problem of correcting a complete correlation matrix for the restricted group, $r(a + x)$ in the Thorndike notation, to that of the unrestricted group, $R(a + x)$, is broken down into two separate problems, correcting r_{ax} to R_{ax} and correcting r_{xx} to R_{xx}, thus:

$$\begin{array}{c|c|c|} & a & x \\ \hline a & r_{aa} & r_{ax} \\ \hline x & r_{xa} & r_{xx} \\ \hline \end{array} \xrightarrow{\qquad\qquad} \begin{array}{c|c|c|} & a & x \\ \hline a & R_{aa} & R_{ax} \\ \hline x & R_{xa} & R_{xx} \\ \hline \end{array}$$

$$r(a + x) \qquad\qquad\qquad\qquad R(a + x)$$

where it is assumed that R_{aa} and H_a are known. The two formulas for accomplishing these corrections are:

$$\text{(IIM).} \quad R_{ax} = R_{aa}H_a b_{ax} H_x^{-1}$$

$$\text{(IIIM).} \quad R_{xx} = H_x^{-1} (r_{xx} - b'_{xa} r_{ax} + b'_{xa} H_a R_{aa} H_a b_{ax}) H_x^{-1}$$

Where $b_{ax} = r_{aa}^{-1} r_{ax}$, the partial regression weights for predicting each indirectly restricted variable, x, from the directly restricted variables, a; and H_x is a diagonal matrix obtained from the square root of the diagonal of the matrix resulting from the operations in the parentheses, P_{xx}. Equation IIIM may also be written:

$$R_{xx} = H_x^{-1} P_{xx} H_x^{-1} = D_x^{-1/2} P_{xx} D_x^{-1/2}$$

The reason for naming these formulas IIM and IIIM, respectively, is to emphasize their relation to the Thorndike univariate correction formulas, II and III, respectively. Considering formula IIIM for one directly selected variable and two indirectly selected variables, $R_{aa} = R_{33} = 1$, $H_a = H_3 = \Sigma 3/\sigma 3$ and $b_{ax} = r_{ax} =$

$$\begin{array}{c|c|c|} & 1 & 2 \\ \hline 3 & r_{13} & r_{23} \\ \hline \end{array}$$

Hence formula IIIM becomes:

$$R_{12} = H_x^{-1}\left[\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 1 & r_{12} \\ 2 & r_{12} & 1 \end{array} - \begin{array}{c|c} & 3 \\ \hline 1 & r_{13} \\ 2 & r_{23} \end{array}\ \begin{array}{c|cc} & 1 & 2 \\ \hline 3 & r_{13} & r_{23} \end{array} + \begin{array}{c|c} & 3 \\ \hline 1 & r_{13} \\ 2 & r_{23} \end{array}\ \begin{array}{c|cc} & & \\ \hline 3 & r_{13} & r_{23} \end{array}\ \Sigma_3^2/\sigma_3^2\right] H_x^{-1}$$

$$P_{12} = \begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 1 & r_{12} \\ 2 & r_{12} & 1 \end{array} + \left[\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & r_{13}^2 & r_{13}r_{23} \\ 2 & r_{13}r_{23} & r_{23}^2 \end{array}\right]\left[\Sigma_3^2/\sigma_3^2 - 1\right]$$

$$P_{12} = \begin{vmatrix} 1 + r_{13}^2\left(\Sigma_3^2/\sigma_3^2 - 1\right) & 1 + r_{13}r_{23}\left(\Sigma_3^2/\sigma_3^2 - 1\right) \\ 1 + r_{13}r_{23}\left(\Sigma_3^2/\sigma_3^2 - 1\right) & 1 + r_{23}^2\left(\Sigma_3^2/\sigma_3^2 - 1\right) \end{vmatrix}$$

and,

$$R_{12} = D^{-1/2}P_{12}D^{-1/2} = \begin{vmatrix} 1 & \text{Thorndike univariate correction formula III} \\ \text{Thorndike univariate correction formula III} & 1 \end{vmatrix}$$

Similarly, for a single directly restricted variable and a single indirectly restricted variable, formula IIM becomes:

$$R_{ax} = r_{ax}\,\Sigma_a/\sigma_a\, H_x^{-1} = \frac{r_{ax}\,\Sigma_a/\sigma_a}{\sqrt{1 - r_{ax}^2 + r_{ax}^2\,\Sigma_a^2/\sigma_a^2}}\ ,$$

or Thorndike Univariate Correction Formula II.

Thus, it is seen that formulas IIM and IIIM are generalizations of formulas II and III, respectively, for handling many coefficients at once. Further inspection of the matric formulas reveals that, with multivariate selection, the correlations among directly restricted variables is taken into account. R_{aa} reduces to a 1 x 1 matrix of unity in the univariate selection case. Hence, it may be expected that serial application of univariate selection formulas to correct for multivariate selection will be fallacious since R_{aa} is thereby assumed to be an identity matrix. It will, indeed, be a rare case, where multiple cutoffs involving uncorrelated variables will be used. It should also be noted that it is the correlations among directly restricted variables in the unselected group that is involved here. Hence, it is also fallacious to ignore these correlations simply because $r_{aa} = 1$. The off-diagonals may have been reduced to zero by the selection process itself.

As further clarification and interpretation of the matric fomulations, the formulas given by Gulliksen in terms of covariances were translated to those given by Thorndike in terms of correlations. Starting with equation 38, (Theory of Mental Tests, p. 165) and transposing both sides:

① $$C_{xy} = C'_{xx} w_{xy}$$

but $$C_{xy} = V_x^{1/2} R_{xy} V_y^{1/2},$$

$$C_{xx} = V_x^{1/2} R_{xx} V_x^{1/2}, \text{ and}$$

$$w_{xy} = v_x^{-1/2} b_{xy} v_y^{1/2},$$

Substituting in ① gives:

② $$V_x^{1/2} R_{xy} V_y^{1/2} = \left[V_x^{1/2} R_{xx} V_x^{1/2}\right] \left[v_x^{-1/2} b_{xy} v_y^{1/2}\right]$$

Premultiplying both sides by $V_x^{-1/2}$ and postmultiplying both sides by $V_y^{-1/2}$ gives:

③ $$R_{xy} = R_{xx} V_x^{1/2} v_x^{-1/2} b_{xy} v_y^{1/2} V_y^{-1/2}$$

but $$V_x^{1/2} v_x^{-1/2} = H_x \text{ and } V_y^{1/2} v_y^{-1/2} = H_y$$

Hence ③ becomes:

④ $$R_{xy} = R_{xx} H_x b_{xy} H_y^{-1}.$$

By translating subscripts to Thorndike notation ④ becomes:

⑤ $$R_{ax} = R_{aa} H_a b_{ax} H_x^{-1}$$

which is identical with formula IIm.

If one starts with equation #42 (Theory of Mental Tests, p. 166):

⑥ $$C_{yy} = c_{yy} + w'_{yx} (C_{xy} - c_{xy}),$$

and similar substitutions are made to convert covariance matrices to correlation matrices, ⑥ becomes:

⑦ $$V_y^{1/2} R_{yy} V_y^{1/2} = v_y^{1/2} r_{yy} v_y^{1/2} + v_y^{1/2} b'_{yx} v_x^{-1/2} \Big[V_x^{1/2} R_{xy} V_y^{1/2} - v_x^{1/2} r_{xy} v_y^{1/}$$

Pre- and post-multiplying both sides by $V_y^{-1/2}$, and rearranging:

⑧ $$R_{yy} = \Big[V_y^{-1/2} v_y^{1/2} r_{yy} v_y^{1/2} V_y^{-1/2} \Big] - \Big[V_y^{-1/2} v_y^{1/2} v_x^{-1/2} b'_{yx} v_x^{1/2} r_{xy} v_y^{1/2} V_y^{-1/}$$
$$+ \Big[V_y^{-1/2} v_y^{1/2} v_x^{-1/2} b'_{yx} V_x^{1/2} R_{xy} V_y^{1/2} V_y^{-1/2} \Big]$$

but $V_y^{-1/2} v_y^{1/2} = v_y^{1/2} V_y^{-1/2} = H_y^{-1}$, and:

⑨ $$R_{yy} = H_y^{-1} r_{yy} H_y^{-1} - H_y^{-1} b'_{yx} r_{xy} H_y^{-1} + H_y^{-1} \left(v_x^{-1/2} b'_{yx} V_x^{1/2} \right) R_{xy} I$$

However, $v_x^{-1/2} b'_{yx} V_x^{1/2} = b'_{yx} H_x$ and from ④, $R_{xy} = R_{xx} H_x b_{xy} H_y^{-1}$. Substituting in ⑨ and factoring H_y^{-1} yields:

⑩ $$R_{yy} = H_y^{-1} \Big[r_{yy} - b'_{yx} r_{xy} + b'_{yx} H_x R_{xx} H_x b_{xy} \Big] H_y^{-1}$$

which, when changed to Thorndike notation, reads:

⑪ $$R_{xx} = H_x^{-1} \Big[r_{xx} - b'_{xa} r_{ax} + b'_{xa} H_a R_{aa} H_a b_{ax} \Big] H_x^{-1}$$

or formula IIIM.

Attention may now be focused on the empirical evaluation of these matrix formulations. These studies were carried out using an experimental population previously prepared, using punched card procedures, and described in the first memorandum of this series (Studies in Methodology - I. Description of an Experimental Domain for Methodological Studies). The theoretical correlations and distribution statistics for the unrestricted population are given in Table 2. A random sample of 500 cases was obtained, and then doubly restricted by first selecting the 300 cases having the highest scores on the composite test, #11, and then selecting the 180 cases, from the 300 case sample, having the highest scores on test #1. The intercorrelations and distribution statistics for this doubly restricted sample are shown in Table 3.

To demonstrate the efficacy of formulas IIM and IIIM, the intercorrelations for variables 1, 3, 5, 6, 10, and 11 were used. Restriction is on variables 1 and 11. Variables 3 and 5 are weighted into the restricting composite test, #11, and variables 6 and 10 are not so weighted. Table 4 shows the given matrix, the corrected matrix, the "true" population matrix, and the error matrix, the latter being obtained by subtracting the "true" population matrix from the corrected matrix. Corrections were carried out using $R_{1:11} = .444$. The magnitudes of the errors are attributable to the sampling errors in the given correlation matrix ($N = 180$; $\sigma_{r_0} = \frac{1}{\sqrt{179}} = 0.075$). It is apparent that the linear dependence variables 3, 5, and 11 have not distorted the correction process. However, tests 3 and 5 are implicitly selected variables. The effect of linear dependence among *explicitly* selected variables is neither known nor likely to be encountered.

The efficacy of the matrix formulas for the special case of univariate selection is easily demonstrated. The sample of 500 cases was subjected to *single* restriction by taking the 300 cases with highest scores on composite test #11.

The intercorrelations for variables 1, 3, 5, 6 and 11 were corrected by the matrix formulas and compared with the corrections obtained for each coefficient by formulas II or III. The resulting error matrices were identical.

Although the formidable appearance of the matrix formulas has probably discouraged their wider use, the frequency with which correlation matrices from restricted groups are encountered in Air Force data would seem to justify more frequent use of these formulas. They *must* be used when selection is multivariate and the restriction variables are correlated. When selection is univariate, the procedure is highly efficient and requires less time than correcting each coefficient separately. It is rare that selection will have occurred on more than two variables and hence, Thorndike's statement about the laborious nature of the computations "when several variables are directly restricted" (*Personnel Selection*, p. 176), while true, need not discourage their use for the more common univariate and bivariate selection problems.

TABLE 2

Theoretical Intercorrelations* and Distribution Statistics for Experimental Domain (N → ∞)

Test	1	2	3	4	5	6	7	8	9	10	mean	S.D.
1											8.100	1.96
2	000										8.160	1.96
3	640	400									8.160	1.96
4	560	000	560								8.550	1.96
5	000	320	200	350							9.680	1.96
6	240	480	540	410	520						10.870	1.96
7	400	560	750	350	280	570					9.345	1.96
8	240	000	240	610	560	410	150				8.100	1.96
9	000	240	150	300	540	420	210	480			9.680	1.96
10	320	320	520	480	440	520	480	440	360		11.100	1.96
**11	444	346	720	693	741	670	580	763	495	639	85.980	14.488

*Decimal points have been omitted

**Test 11 is a raw score composite = 4 x (Test 3) + 3 x (Test 5) + 3 x (Test 8)

TABLE 3

Intercorrelations* and Distribution Statistics for the Doubly Restricted Sample (N = 180)**

Test	1	2	3	4	5	6	7	8	9	10	Mean	S.D.
1											10.011	1.134
2	-061										8.202	1.757
3	403	445									9.753	1.574
4	254	-137	253								10.030	1.558
5	-156	134	-090	142							10.229	1.513
6	-014	313	337	138	328						11.626	1.700
7	196	544	676	106	017	332					10.468	1.746
8	040	-213	-189	493	392	131	-239				8.911	1.600
9	-087	116	019	163	335	330	055	322			9.977	1.724
10	166	265	391	138	261	395	389	153	220		12.104	1.724
***11	220	260	544	502	640	463	347	589	322	488	96.433	9.156

*Decimal points have been omitted

**Double restriction has occurred on tests #1 and #11.

***Test #11 is a raw score composite = 4.0 (Test 3) + 3.0 (Test 5) + 3.0 (Test 8).

TABLE 4

Efficacy of Matrix Formulation for Correcting for Double Restriction of Range*

Given $r_{(a+x)}$

	3	5	6	10	1	11
3	1000					
5	-090	1000				
6	337	328	1000			
10	391	281	395	1000		
1	403	-156	-014	166	1000	
11	544	640	460	488	220	1000

Corrected $R_{(a+x)}$

	3	5	6	10	1	11
3	1000					
5	180	1000				
6	465	497	1000			
10	590	443	525	1000		
1	634	-033	119	371	1000	
11	734	718	605	673	(444)**	1000

"True" $R_{(a+x)}$

	3	5	6	10	1	11
3	1000					
5	200	1000				
6	540	520	1000			
10	520	440	520	1000		
1	640	000	240	320	1000	
11	720	741	670	639	444	1000

$E_{(a+x)}$

	3	5	6	10	1	11
3	000					
5	-020	000				
6	-075	-023	000			
10	070	003	005	000		
1	-006	-033	-121	051	000	
11	014	-023	-065	034	(000)	000

$$\sigma r_o = \frac{1}{\sqrt{179}} = 0.075$$

*Restriction has occurred on Variables 1 and 11. Decimal points have been omitted

**$R_{1:11}$ given rather than obtained by correction

4-18

PERSONNEL RESEARCH LABORATORY
AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER
AIR RESEARCH AND DEVELOPMENT COMMAND
LACKLAND AIR FORCE BASE
San Antonio, Texas

STAFF RESEARCH MEMORANDUM*
Task 77006

12 April 1954

Studies in Methodology
V. The Efficacy of Two Variants of Thorndike Formula #7 for Correcting Correlation Coefficients for Range Restriction

John A. Creager

The efficacy of the three basic univariate formulas for correcting correlation coefficients for range restriction was discussed in a previous Staff Research Memorandum. Thorndike correction Formula #7 (Thorndike, R. L. Personnel Selection, p. 174) is applicable for correcting the coefficient of correlation between two variables when direct restriction has occurred on a third variable. This formula requires knowledge of three correlations obtained for the restricted population: r_{12}, the coefficient being corrected; r_{13} and r_{23}, the correlations of each variable with the directly restricted variable. In certain instances, the correlations with the directly restricted variable may be known only for the unrestricted group. Thorndike gives a variant of Formula #7 (#8) for the situation where one of the correlations involving the directly restricted variable is known for the restricted group and the other is known for the unrestricted group. It is the purpose of this memorandum to report a small study carried out to:

a. show how Thorndike's Formula #8 was derived.

b. derive a variant of Formula #7 where both correlations involving the directly restricted variable are known only for the unrestricted group.

c. demonstrate the efficacy of both formulas for obtaining an estimate of the unrestricted value, R_{12}.

The need for these two variants of Thorndike's Formula #7, while not common, can arise in practical situations. Thus, Formula #8 would be used in a validation study where the correlation between a test whose validity is under investigation and the selection test is known only for the unselected group. If, in addition, the unrestricted validity of the selection score must also be used, a second variant of Thorndike's Formula #7 would be needed.

---

*AFPTRC Staff Research Memoranda are informal papers intended to record opinions and preliminary reports of studies. They may be expanded, modified, or withdrawn at any time and hence are not suitable for inclusion or reference in more permanent reports of a scientific or technical character.

In line with the previous studies in this series on range restriction formulas, it will be convenient to refer to Thorndike's Formula #7 as univariate correction Formula III, Thorndike's Formula #8 as univariate correction Formula III A, and the second variant to be derived as univariate correction Formula III B.

If test 1 is the test under investigation, test 2 a criterion variable, and test 3 the selection test, the basic univariate correction Formula III reads as follows:

$$R_{12} = \frac{r_{12} + r_{13}r_{23}\left(\frac{S_3^2}{s_3^2} - 1\right)}{\sqrt{\left[1 + r_{13}^2\left(\frac{S_3^2}{s_3^2} - 1\right)\right]\left[1 + r_{23}^2\left(\frac{S_3^2}{s_3^2} - 1\right)\right]}} \tag{1}$$

where S_3 is the standard deviation of the unrestricted group and s_3 that of the restricted group.

The derivations of the variant formulas involve substituting expressions for r_{13} and r_{23} in formula (1). These expressions are obtained by writing univariate correction Formula II for R_{13} and R_{23}, squaring, and solving for r_{13}^2 and r_{23}^2, respectively:

$$R_{13} = \frac{r_{13}\,\frac{S_3}{s_3}}{\sqrt{1 - r_{13}^2 + r_{13}^2\,\frac{S_3^2}{s_3^2}}} \tag{2}$$

$$R_{23} = \frac{r_{23}\,\frac{S_3}{s_3}}{\sqrt{1 - r_{23}^2 + r_{13}^2\,\frac{S_3^2}{s_3^2}}} \tag{3}$$

$$r_{13}^2 = \frac{R_{13}^2 \dfrac{s_3^2}{S_3^2}}{1 + R_{13}^2 \left(\dfrac{s_3^2}{S_3^2} - 1\right)} \tag{4}$$

$$r_{23}^2 = \frac{R_{23}^2 \dfrac{s_3^2}{S_3^2}}{1 + R_{23}^2 \left(\dfrac{s_3^2}{S_3^2} - 1\right)} \tag{5}$$

Substituting (4) in (1) gives:

$$R_{12} = \frac{r_{12} + \left[\dfrac{R_{13} \dfrac{s_3}{S_3}}{\sqrt{1 + R_{13}^2 \left(\dfrac{s_3^2}{S_3^2} - 1\right)}}\right] \cdot r_{23} \left(\dfrac{S_3^2}{s_3^2} - 1\right)}{\sqrt{1 + \dfrac{R_{13}^2 \left(\dfrac{S_3^2}{s_3^2} - 1\right)}{R_{13}^2 \left(1 - \dfrac{S_3^2}{s_3^2}\right) + \dfrac{S_3^2}{s_3^2}}} \; \sqrt{1 + r_{23}^2 \left(\dfrac{S_3^2}{s_3^2} - 1\right)}} \tag{6}$$

which may be simplified to:

$$R_{12} = \frac{r_{12} \sqrt{1 + R_{13}^2 \left[\dfrac{s_3^2}{S_3^2} - 1\right]} + R_{13}\, r_{23} \left[\dfrac{S_3}{s_3} - \dfrac{s_3}{S_3}\right]}{\sqrt{1 + r_{23}^2 \left(\dfrac{S_3^2}{s_3^2} - 1\right)}} \tag{7}$$

, Formula III A

This is identical to Throndike's Formula #8.

Substituting (5) in (7) gives:

$$R_{12} = \frac{r_{12}\sqrt{1 + R_{13}^2\left(\frac{s_3^2}{S_3^2} - 1\right)} + R_{13}\left[\frac{R_{23}\,\frac{s_3}{S_3}}{\sqrt{1 + R_{23}^2\left(\frac{s_3^2}{S_3^2} - 1\right)}}\right]\left[\frac{S_3}{s_3} - \frac{s_3}{S_3}\right]}{\sqrt{1 + \frac{R_{23}^2\,\frac{s_3^2}{S_3^2}}{1 + R_{23}^2\left[\frac{s_3^2}{S_3^2} - 1\right]}\left[\frac{S_3^2}{s_3^2} - 1\right]}} \tag{8}$$

which may be simplified to:

$$R_{12} = r_{12}\sqrt{1 + R_{13}^2\left[\frac{s_3^2}{S_3^2} - 1\right]}\sqrt{1 + R_{23}^2\left[\frac{s_3^2}{S_3^2} - 1\right]} - R_{13}\,R_{23}\left[\frac{s_3^2}{S_3^2} - 1\right], \tag{9}$$

which is Formula III B.

The intercorrelation matrix for the unrestricted population is shown in Table 1. The intercorrelation matrix for Restricted Sample I is shown in Table 2. This matrix was corrected by univariate correction Formulas III A and III B, resulting in the intercorrelations in Table 3 and the errors in Table 4. The upper half of each matrix refers to those correlations corrected by III A; the lower half, those corrected by III B. The errors are of the same order of magnitude as those for the basic univariate correction Formula III, and can be attributed to sampling errors in the original restricted sample. ($n = 300$; $\sigma_{r_0} = .058$)

Theoretical Intercorrelations* and Distribution Statistics for Experimental Domain $(N \to \infty)$

Test	1	2	3	4	5	6	7	8	9	10	Mean	S.D.
1											8.100	1.96
2	000										8.160	1.96
3	640	400									8.160	1.96
4	560	000	560								8.550	1.96
5	000	320	200	350							9.680	1.96
6	240	480	540	410	520						10.870	1.96
7	400	560	750	350	280	570					9.345	1.96
8	240	000	240	610	560	410	150				8.100	1.96
9	000	240	150	300	540	420	210	480			9.680	1.96
10	320	320	520	480	440	520	480	440	360		11.100	1.96
**11	444	346	720	693	741	670	580	763	495	639	85.980	14.488

---

*Decimal points have been omitted
**Test 11 is a raw score composite = 4 x (Test 3) + 3 x (Test 5) + 3 x (Test 8)

Intercorrelations* and Distribution Statistics for Restricted Sample #1 (N = 300)**

Test	1	2	3	4	5	6	7	8	9	10	Mean	S.D.
1											8.744	1.897
2	-214										8.542	1.814
3	575	293									9.159	1.666
4	470	-184	405								9.430	1.712
5	-306	213	-219	028							10.518	1.548
6	006	346	319	198	294						11.596	1.673
7	330	460	673	224	-042	374					10.103	1.737
8	007	-155	-138	369	335	183	-149				8.910	1.560
9	-172	121	-084	096	396	293	004	359			10.186	1.783
10	152	240	410	239	216	412	397	157	178		11.988	1.777
***11	272	247	559	508	530	485	401	593	330	499	94.915	8.962

---

*Decimal points have been omitted

**Restriction has occurred on Test #11.

***Test 11 is a raw score composite = 4.0 (Test 3) + 3.0 (Test 5) + 3.0 (Test 8)

Intercorrelations for Restricted Sample I Corrected by Formulas III A & III B*

Test	1	2	3	4	5	6	7	8	9	10
1		-106	632	550	-045	177	426	201	-023	287
2	-098		401	016	340	438	528	061	222	349
3	641	386		593	192	526	753	258	162	581
4	559	000	588		335	427	414	573	285	446
5	-030	324	182	336		501	224	557	520	436
6	189	426	520	428	509		523	439	434	566
7	434	519	752	415	235	523		166	180	534
8	215	044	249	575	567	440	167		498	410
9	-012	212	157	286	523	434	181	498		336
10	298	336	577	447	444	567	534	410	337	

---

*Decimal points have been omitted. The upper half of the matrix was corrected by Formula III A; the lower half by Formula III B.

TABLE 4

Error Matrices for Formulas III A & III B*

Test	1	2	3	4	5	6	7	8	9	10
1		-106	-008	-010	-045	-063	026	-039	-023	-033
2	-098		001	016	020	-042	-032	061	-018	029
3	001	-014		033	-008	-014	003	018	012	061
4	-001	000	028		-015	017	064	-037	-015	-034
5	-030	004	-018	-014		-019	-056	-003	-020	-004
6	-051	-054	-020	018	-011		-047	-029	014	046
7	034	-041	002	065	-045	-047		016	-030	054
8	-025	044	009	-035	007	-030	017		018	-030
9	-012	-028	007	-014	-017	014	-029	018		-024
10	-022	016	057	-033	004	047	054	-030	-023	

---

*Decimal points have been omitted. The upper half of the matrix was corrected by Formula III A; the lower half by Formula III B.

Personnel Laboratory
Wright Air Development Division
Air Research and Development Command
United States Air Force
Lackland Air Force Base, Texas

Technical Memorandum
WWRDP-TM-60-40
27 September 1960

Selection and Classification Branch
Project 7717-87003

ON THE USE OF A COMPOSITE SIMULATING COMPLEX SELECTION

Real problems are seldom as simple, clear-cut, and neatly soluble as a graduate student might expect from perusal of his textbooks. Consider, for example, a situation where selection for admission to a training program involves:

1. A selection composite consisting of two aptitude test composites, some demographic variables, a special ability test score, a personality test score, and a rating of past performance.

2. Elimination of those not meeting minimum cutoffs on one of the aptitude composites, one of the demographic variables, and the special ability test.

3. Application of the selection composite to the group meeting the multiple cutoff requirements, except that bonus points are added to the composite scores of certain candidates for various extraneous factors.

4. Elimination of 10% of the selectees by administrative action, negatively correlated with other factors in the system, and based in part on an interview of the candidate.

Research under such conditions may be somewhat tenuous, even where the worker has great insight into the system and statistical sophistication. Evaluating the selection or components thereof, introducing controls in training studies, or correcting validities for range restriction are rather formidable, and may involve so many tenuous assumptions and devious practices as to cast doubt on the results.

This memorandum proposes such a complex selection process be simulated by creating a multiple regression system, generated on the full applicant group. This system uses as criterion a dichotomous variable, "1" if the candidate was ultimately selected, "0" if rejected. Scores on the selection variables may be used as predictors. If this multiple correlation is reasonably high (as it usually would be), the regression composite may be taken as simulating the complex selection. A high correlation indicates that most of the contributing variables of the actual system (or their equivalent) have been taken into account. Also, by examination of the effective weights in the simulation composite, useful information may be obtained regarding the role of the various selection variables.

If the multiple correlation is low, either important aspects of the selection have not been taken into account and further investigation of the bases for selection are indicated; or, the selection is not being carried out in accordance with the explicitly stated rules.

If the distribution of scores on the simulated selection composite is cut at the actual selection ratio, the phi coefficient between actual and "predicted" selection may be regarded as an additional index of simulation. It should be noted that phi may equal unity even when the multiple correlation is appreciably less than 1. If there is perfect (or near perfect) accounting of actual selection as measured by the phi coefficient, the simulation composite may be considered as a simpler selection device, accomplishing the same result as the elaborate and complex procedures actually used. For this purpose the simulation does not have to be perfect as measured by the multiple correlation. Actual recommendation of the simulation composite in lieu of the complex selection procedure would assume no change in either the intended bases of selection or the selection ratio. If such changes are contemplated, the appropriate simulation composite and selection ratio can be examined for the consequences.

The level of the multiple correlation, and hence degree of simulation, may be increased by introducing dichotomous predictor variables based on the level of the multiple cutoffs in the system. Dichotomous variables are also indicated where arbitrary metric weighting has been used for various levels of an ordered qualitative variate (e.g. military rank). Introduction of apparently extraneous factors may also increase simulation and provide further information on the selection process.

The simulation composite may also be useful as a basis for range restriction corrections where the multivariate methods would not be feasible. It would not be necessary to assume that selection was confined to truncations in a multinormal applicant distribution. However, for this purpose, the validity of the simulated selection must be very high as measured by the multiple (probably greater than .90). This procedure tends to undercorrect rather than overcorrect.

An initial tryout of the simulation method was performed by Mr. Valentine, using some OCS data available on applicants prescreened at 5 on Officer Quality. The qualified applicant group was then subjected to complex selection in accordance with rules that were operational at that time (but which have since been modified). No attempt was made to introduce some of the refinements in the simulation regression as suggested above, e.g., level dichotomies. The multiple with actual selection was .70. Regressed selection scores were computed on a random sample of 90 cases, stratified by selection-rejection so as to preserve the initial selection ratio. The regressed score distribution was cut at the selection ratio (.52 selected) and the phi obtained between "predicted" and actual selection was .91. Of the 90 cases, two selectees were "predicted" as rejectees, and two rejectees "predicted" to be selectees, a total of four errors of classification in 90 decisions.

Prepared by:
John A. Creager, WWRDPS

PUBLICATION REVIEW

This report has been reviewed and is approved.

A. Carp

A. Carp, Technical Director
Personnel Laboratory

Distribution: WADD

Personnel Laboratory
Wright Air Development Division
Air Research and Development Command
United States Air Force
Lackland Air Force Base, Texas

Technical Memorandum
WWRDP-TM-60-40
27 September 1960

Selection and Classification Branch
Project 7717-87003

ON THE USE OF A COMPOSITE SIMULATING COMPLEX SELECTION

Real problems are seldom as simple, clear-cut, and neatly soluble as a graduate student might expect from perusal of his textbooks. Consider, for example, a situation where selection for admission to a training program involves:

1. A selection composite consisting of two aptitude test composites, some demographic variables, a special ability test score, a personality test score, and a rating of past performance.

2. Elimination of those not meeting minimum cutoffs on one of the aptitude composites, one of the demographic variables, and the special ability test.

3. Application of the selection composite to the group meeting the multiple cutoff requirements, except that bonus points are added to the composite scores of certain candidates for various extraneous factors.

4. Elimination of 10% of the selectees by administrative action, negatively correlated with other factors in the system, and based in part on an interview of the candidate.

Research under such conditions may be somewhat tenuous, even where the worker has great insight into the system and statistical sophistication. Evaluating the selection or components thereof, introducing controls in training studies, or correcting validities for range restriction are rather formidable, and may involve so many tenuous assumptions and devious practices as to cast doubt on the results.

This memorandum proposes such a complex selection process be simulated by creating a multiple regression system, generated on the full applicant group. This system uses as criterion a dichotomous variable, "1" if the candidate was ultimately selected, "0" if rejected. Scores on the selection variables may be used as predictors. If this multiple correlation is reasonably high (as it usually would be), the regression composite may be taken as simulating the complex selection. A high correlation indicates that most of the contributing variables of the actual system (or their equivalent) have been taken into account. Also, by examination of the effective weights in the simulation composite, useful information may be obtained regarding the role of the various selection variables.

If the multiple correlation is low, either important aspects of the selection have not been taken into account and further investigation of the bases for selection are indicated; or, the selection is not being carried out in accordance with the explicitly stated rules.

If the distribution of scores on the simulated selection composite is cut at the actual selection ratio, the phi coefficient between actual and "predicted" selection may be regarded as an additional index of simulation. It should be noted that phi may equal unity even when the multiple correlation is appreciably less than 1. If there is perfect (or near perfect) accounting of actual selection as measured by the phi coefficient, the simulation composite may be considered as a simpler selection device, accomplishing the same result as the elaborate and complex procedures actually used. For this purpose the simulation does not have to be perfect as measured by the multiple correlation. Actual recommendation of the simulation composite in lieu of the complex selection procedure would assume no change in either the intended bases of selection or the selection ratio. If such changes are contemplated, the appropriate simulation composite and selection ratio can be examined for the consequences.

The level of the multiple correlation, and hence degree of simulation, may be increased by introducing dichotomous predictor variables based on the level of the multiple cutoffs in the system. Dichotomous variables are also indicated where arbitrary metric weighting has been used for various levels of an ordered qualitative variate (e.g. military rank). Introduction of apparently extraneous factors may also increase simulation and provide further information on the selection process.

The simulation composite may also be useful as a basis for range restriction corrections where the multivariate methods would not be feasible. It would not be necessary to assume that selection was confined to truncations in a multinormal applicant distribution. However, for this purpose, the validity of the simulated selection must be very high as measured by the multiple (probably greater than .90). This procedure tends to undercorrect rather than overcorrect.

An initial tryout of the simulation method was performed by Mr. Valentine, using some OCS data available on applicants prescreened at 5 on Officer Quality. The qualified applicant group was then subjected to complex selection in accordance with rules that were operational at that time (but which have since been modified). No attempt was made to introduce some of the refinements in the simulation regression as suggested above, e.g., level dichotomies. The multiple with actual selection was .70. Regressed selection scores were computed on a random sample of 90 cases, stratified by selection-rejection so as to preserve the initial selection ratio. The regressed score distribution was cut at the selection ratio (.52 selected) and the phi obtained between "predicted" and actual selection

PERSONNEL RESEARCH LABORATORY
Air Force Personnel and Training Research Center
Air Research and Development Command
Lackland Air Force Base, Texas

LABORATORY NOTE PRL-LN-55-13[1]
7701-77023

26 April 1955

Correcting Correlation Coefficients for Selection When the Nature of the Selection is Unknown

John A. Creager
Robert G. Smith

Problem

In many practical problems sources of selection in addition to direct truncation render the Thorndike correction formulas inadequate for estimating correlation coefficients for an unrestricted population. This note presents a procedure designed to correct coefficients attenuated by selection, without making the highly restrictive assumptions usually more or less violated in applying the Thorndike formulas.

Assumptions

The method presented in this note assumes that:

1. the unrestricted bivariate frequency distribution is normal and homoscedastic for both variables,

2. selection has resulted only in decreased frequencies in certain cells of the bivariate frequency distribution,

3. the selection ratio is known.

The first assumption implies that the unrestricted regressions are linear. The second assumption rules out additions to the sample due to transfers, holdovers, etc.

The last assumption ideally refers to total per cent losses from testing to criterion data collection regardless of source. In practice the selection will generally include truncation, as qualified by cases admitted below the cut-off to fulfill quotas, administrative losses above the cut-off, early eliminations, etc. Thus there is no restriction of selection to truncation of the tail of a distribution or assumption that the slope of one regression line be unaffected by the selection.

---

[1] This paper is an informal note and is subject to modification or withdrawal at any time. If referenced, it should be described as an "unpublished draft."

Method

A normal bivariate scatter-plot for a correlation of .294 on 1,000 cases is presented in Table 1. This was subjected to truncation and several arbitrary losses throughout the matrix to yield the restricted scatter-plot in Table 2. This is bordered by the marginal frequency distributions in the restricted sample (N=550). In a practical problem one obtains this matrix and the per cent loss (45.0). The problem is then to try to reproduce the normal bivariate frequency distribution from which the selection sample was obtained. From the per cent loss and restricted sample size it may be inferred that the unrestricted sample size was 1,000. The marginal frequencies may then be determined from the areas under the normal curve. In this example stanine distributions were used. The scatter-plot in Table 2 is then further bordered by a row and column of discrepancies (d-values) between the row (or column) sum and the unrestricted marginal frequencies.

Table 1

Normal Bivariate Frequency Distribution
(N = 1000; r = .294)

	1	2	3	4	5	6	7	8	9	
9	0	1	2	4	7	8	8	6	5	
8	1	2	4	8	12	14	11	7	6	
7	2	4	10	18	24	24	19	11	8	
6	4	8	18	29	36	34	24	14	8	
5	7	12	24	36	41	36	24	12	7	
4	8	14	24	34	36	29	18	8	4	
3	8	11	19	24	24	18	10	4	2	
2	6	7	11	14	12	8	4	2	1	
1	5	6	8	8	7	4	2	1	0	
Σ_1	41	65	120	175	199	175	120	65	41	$\Sigma\Sigma$ = 1001
*Σ_2	40	66	121	174	198	174	121	66	40	$\Sigma\Sigma$ = 1000

*Theoretical sum used as basis of d-values.

Table 2

Restricted Bivariate Frequency Distribution
(N = 550; r = .186)

	1	2	3	4	5	6	7	8	9	Σ	d	Check Σ
9	0	0	0	0	7	7	7	5	5	31	9	40
8	0	0	0	5	7	12	10	5	6	45	21	66
7	0	0	0	15	20	24	15	10	6	90	31	121
6	0	0	0	10	30	30	20	10	5	105	69	174
5	0	0	20	35	0	20	20	10	5	110	88	198
4	0	0	0	0	20	25	10	7	2	64	110	174
3	0	0	0	10	20	15	10	4	0	59	62	121
2	0	0	0	0	10	8	4	2	1	25	41	66
1	0	0	0	8	7	3	2	1	0	21	19	40
Σ	0	0	20	83	121	144	98	54	30	550	—	—
d	40	66	101	91	77	30	23	12	10	—	450	—
Check Σ	40	66	121	174	198	174	121	66	40	—	—	1000

Noting that the bivariate normal distribution is diagonally symmetrical about the center cell ($\overline{xy}$), the frequencies in the restricted bivariate distribution are built up to yield a symmetrical distribution. For example, cell 3, 3; 3, 6; 4, 7; and 7, 7 should contain the same frequency. The largest value, 15 (in cell 7, 7), is placed in all four cells. This is done for the whole matrix until the desired symmetry is obtained with a minimal addition of cases. The row and column sums, and the d-values are readjusted. The resulting matrix is shown in Table 3.

Noting that each row and column of a bivariate distribution is unimodal, one next proceeds to remove any inversions by increasing the frequency in the "troublesome" cell to the lowest value in an adjacent cell. When this is done for the whole matrix, symmetry will be retained, inversions in the marginals will disappear, but will now appear among the d-values. The resulting matrix for the example is shown in Table 4 with marginal and adjusted d-values.

Table 3

Symmetrized Bivariate Frequency Distribution
(r = .333)

	1	2	3	4	5	6	7	8	9	Σ	d	Check Σ
9	0	1	0	2	7	7	7	6	5	35	5	40
8	1	2	4	7	10	12	10	5	6	57	9	66
7	0	4	10	15	20	24	15	10	7	105	16	121
6	2	7	15	20	35	30	24	12	7	152	22	174
5	7	10	20	35	0	35	20	10	7	144	54	198
4	7	12	24	30	35	20	15	7	2	152	22	174
3	7	10	15	24	20	15	10	4	0	105	16	121
2	6	5	10	12	10	7	4	2	1	57	9	66
1	5	6	7	7	7	2	0	1	0	35	5	40
Σ	35	57	105	152	144	152	105	57	35	842	—	—
d	5	9	16	22	54	22	16	9	5	—	158	—
Check Σ	40	66	121	174	198	174	121	66	40	—	—	1000

Table 4

Bivariate Frequency Distribution After Removal of
Inversions in the Arrays (r = .321)

	1	2	3	4	5	6	7	8	9	Σ	d	Check Σ
9	0	1	1	2	7	7	7	6	5	36	4	40
8	1	2	4	7	10	12	10	6	6	58	8	66
7	1	4	10	15	24	24	15	10	7	106	15	121
6	2	7	15	20	35	30	24	12	7	152	22	174
5	7	10	20	35	35	35	20	10	7	187	11	198
4	7	12	24	30	35	20	15	7	2	152	22	174
3	7	10	15	24	24	15	10	4	1	106	15	121
2	6	6	10	12	10	7	4	2	1	58	8	66
1	5	6	7	7	7	2	1	1	0	36	4	40
Σ	36	58	106	152	187	152	106	58	36	891	—	—
d	4	8	15	22	11	22	15	8	4	—	109	—
Check Σ	40	66	121	174	198	174	121	66	40	—	—	1000

The next step is to reduce the d-values by applying "contingency" corrections to the cell frequencies. The correction for a given cell, C_{ij}, is $\frac{d_i d_j}{\Sigma d}$ where Σd is the total of all deviation values for the matrix.

Where the correction is nearly halfway between two integral values, the larger one is taken and recorded with a minus sign after it. The marginal and d-values are readjusted. The result of this operation is shown in Table 5.

Table 5

Bivariate Frequency Distribution After Application of Contingency Corrections (r = .284)

	1	2	3	4	5	6	7	8	9	Σ	d	Check Σ
9	0	1	2	3	7	8	8	6	5	40	0	40
8	1	3-	5	9-	11-	14	11	7	6	67	-1	66
7	2	5	12	18	21	27	17	11	8	121	0	121
6	3	9-	18	25	37	35	27	14	8	176	-2	174
5	7	11-	21	37	36	37	21	11-	7	188	10	198
4	8	14	27	35	37	25	19	9-	3	176	-2	174
3	8	11	17	27	21	18	12	5	2	121	0	121
2	6	7	11	14	11-	9-	5	3-	1	67	-1	66
1	5	6	8	8	7	3	2	1	0	40	0	40
Σ	40	67	121	176	188	176	121	67	40	996	—	—
d	0	-1	0	-2	10	-2	0	-1	0	—	4	—
Check Σ	40	66	121	174	198	174	121	66	40	—	—	1000

Small adjustments are made reducing frequencies by 1 which have values in a row with a negative d-value. One starts with cells most removed from the regression line until minus signs are removed or the d-value for the row is no longer negative, whichever occurs first. Finally the contingency principle is reapplied using readjusted d-values. The finally obtained corrected scatter-plot is shown in Table 6. Table 7 shows the discrepancies between Tables 1 and 6.

Table 6

Bivariate Frequency Distribution Corrected for Arbitrary Selection Losses (r = .293)

	1	2	3	4	5	6	7	8	9	Σ	d	Check Σ
9	0	1	2	3	7	8	8	6	5	40	0	40
8	1	2	5	8	12	14	11	7	6	66	0	66
7	2	5	12	18	21	27	17	11	8	121	0	121
6	3	8	18	25	36	35	27	14	8	174	0	174
5	7	12	21	36	46	36	21	12	7	198	0	198
4	8	14	27	35	36	25	18	8	3	174	0	174
3	8	11	17	27	21	18	12	5	2	121	0	121
2	6	7	11	14	12	8	5	2	1	66	0	66
1	5	6	8	8	7	3	2	1	0	40	0	40
Σ	40	66	121	174	198	174	121	66	40	1000	—	—
d	0	0	0	0	0	0	0	0	0	—	0	—
Check Σ	40	66	121	174	198	174	121	66	40	—	—	1000

Results

The errors shown in Table 7 are quite small and concentrated for the most part near the regression line. The ±1 errors farther out may be attributed to rounding errors in Table 1.

The corrected correlation coefficient computed from the scatter-plot in Table 6 is .293 (as compared with .294 in Table 1). The uncorrected coefficient computed from Table 2 is .186, which corrected by Thorndike Formula 6 becomes .243.

Table 7

Matrix of Discrepancies Between Unrestricted and Corrected Bivariate Frequency Distributions*

	1	2	3	4	5	6	7	8	9	Σ
9	0	0	0	-1	0	0	0	0	0	-1
8	0	0	+1	0	0	0	0	0	0	+1
7	0	+1	+2	0	-3	+3	-2	0	0	+1
6	-1	0	0	-4	0	+1	+3	0	0	-1
5	0	0	-3	0	+5	0	-3	0	0	-1
4	0	0	+3	+1	0	-4	0	0	-1	-1
3	0	0	-2	+3	-3	0	+2	+1	0	+1
2	0	0	0	0	0	0	+1	0	0	+1
1	0	0	0	0	0	-1	0	0	0	-1

$\Sigma\Sigma = -1$

*Cell values are corrected values from Table 6 minus original values from Table 1.

Conclusion

This note presents a method for correcting correlation coefficients for selection. It is designed to have more general applicability than the Thorndike formulas which assume simple truncation (either direct or indirect) as the sole source of bias. The method presented was illustrated by an example involving an unrestricted correlation of about .30 subjected to various kinds of losses. The example was based on 1,000 cases for the unrestricted sample and 45 per cent loss by selection. Further investigation is required to ascertain the scope and limitations of the method, particularly as it is affected by small sample size and extreme percentage losses. Further investigation is also required to determine the adaptability of the method for special problems frequently encountered in practice. These problems include the cases where the unrestricted marginal distributions are not normal, the unrestricted scatter-plots are curvilinear, or an erroneous estimate is made of the selection ratio.

APPENDIX B

MEMORANDA FOR MISOE

March 3, 1972

Dr. William G. Conroy, Jr.
Division of Occupational Education
1017 Main Street
Winchester, Massachusetts 01890

Dear Bill:

This letter comments primarily on OPs#2 and 4. It is useful to approach the differentiations and instrumentation of the IPPI in terms of roles these elements play in the total MISOE. In the case of the input space, the student data are needed to characterize input to computer flow models, studying manpower issues, etc. They are also needed as control variables in analyses of product and impact outcomes from processes, and therefore, should include indicators of pre-process experiences and capabilities relevant to such outcomes. While student data are needed as such for analyses where the student is the analytic unit, aggregate summary data on the students entering particular programs, schools, etc., are required where these are the analysis units. The distinction between local, state, federal, and other capital data for cost-benefit analyses may require some arbitrary decisions, explicitly stated and uniformly applied, when dollar input to a program coming directly from a LEA indirectly comes from state funds, which in turn may have come partially from federal sources. Thus, the identifiability of expenditures by these distinctions may trip over their lack of independence. Identifying by source chains may be helpful.

At some point we should probably have a look at admissions requirements and variations in such requirements across schools giving the "same" programs. This may be more important in the post secondary programs.

Variable selection and instrumentation in input space seem straightforward except for ensuring equivalence of "scores" (e.g., IQ) from different instruments purporting to measure the same thing, and for ensuring acquisition of prior experience data information mentioned above.

It is in the process space that manipulability and the feedback of results of decision making are most relevant. The present delineation of this space (OP#2, Fig. 2) as elaborated by EW in conference seems excellent, as is the explicit provision for obtaining cost data within this space (OP#2, P. 8), especially for the physical factors and for personnel. Also, I gather, that student perceptions of the process belong here under "perceptual" as an exception to the human factors referring to nonstudent personnel.

One basis for classification of physical factors (within either structural or instructional types) would be on their joint occurrence across schools and programs. It should not be necessary to include in analyses of the products and impacts of process two physical factors which nearly always occur together; or, if the jointly occurring factors were ordered variables rather than qualitative conditions, would it be necessary to measure both.

I take it that the breakdown of an instructional event such as Fortran programming course example into blocks and units yields examples of organizational factors and should included information on sequencing.

EW's delineation of organizational factors, a-d (OP#4, P. 3-4) should include a fifth factor: operating rules such as accessibility of physical equipment to the student. An alternative is to identify such "rules of organization" with decisional behaviors under human factors. The confusion arises because we are dealing with role incumbent decisions about organizational factors.

Obtaining both process and cost information for the process space depends very much on what is already documented at the local level, the degree of consistency in such documentation across schools giving similar programs, and the logistic flexibilities or constraints you may encounter in obtaining data on bases that ensure comparability across potential analysis groups.

Student perceptions of process may be picked up by a simple, objective but confidential questionnaire focused on the nature and amount of teacher contact, fast feedback of evaluations of the student's performance, whether the atmosphere permitted the student to resolve perplexities, and EW's suggestion about the student's feeling of some degree of control over the learning situation.

EW's discussion of decision making in the hierarchical arrangement of the process space with higher level decisions constraining decisions and other process factors at the lower level may have some special analysis implications. The presence or absence of such constraints can be indicated by dichotomous variables in regression. It may be that these constraints can be expressed in constraint equations in the case of linear programming models, or as modified transition probabilities in flow models.

The addressing system for process space information appears reasonable and even necessary to the functioning of the total system. The school, program, and block subscripting arrangement is critical for identifying analysis units. In the case of analyses where the students are the units of analysis, it is crucial that his data include the subscript in order to link the process variables to which he is exposed.

You expressed concern about summing capabilities within and across programs. If I interpret the concern correctly, it is an analytic rather than a product space differentiation problem. In my last letter, I commented on some of the pros and cons of weighted summaries versus configural approaches to combining outputs in both product and impact space as "dependent variables" in analysis.

In your suggestion to expand Figure 4 into a 2-way table in terms of geographic space, you may be able to capitalize for the local, regional, and state on the notion of geopolitical stratification suggested in my last letter. In any case you will probably need geographic and occupational migration data in the later development of the impact space along these lines.

The instrumentation of the impact space is difficult and it is here that I am hopeful that contacts with the DOD occupational analysis systems may be helpful for parts of the job, especially in the "Self" portions. Supplementing with followup questionnaires to subjects and their employers should also be helpful in both self and society portions.

Although I was thinking in terms of predicting impact space variables from process (and product) variables, your example of "equal opportunity" measures suggests the juggling of system parameters in simulation and comparison of different racial mixes within occupations that result with both the current actual mix and the "ideal" mixes defined by someone's value judgement.

In the instrumentation task, it will be important for analysis that instrument reliability be high and, in all cases, either known or plausibly estimable. While a good deal of useful summary descriptive information can be obtained efficiently with moderately reliable instruments, measurement error of feeding back erroneous inferences into the system may be cumulative. A forthcoming ACE report discusses these matters and provides a useful list of references; even though our concerns are relevant to higher education, the same principles apply to occupational educational data.

Another instrumentation issue is the possibility, where the same instrument is to be administered to large segments of the sample, of designing or adopting instruments that can be read on an optical scanner should be considered. Given sufficient volume (N of 5,000 or so) a great deal of information can be obtained very efficiently and result in data input tapes for the computerized aspects of the system. This applies to instruments with objective formats (check lists, multiple choice, etc., rather than open-ended or essay response). You will probably have this constraint anyway where decentralized administration and limited testing time are at issue.

I plan to write one more letter commenting on the computerized information system (Tasks 5-8, and CP#3). The requirement that the system be ongoing and expandable suggests that EW's subscripting codes for data units may have to be expanded to identify the time-cohort involved. This may not be necessary if all data are permanently stored by time-cohort on labeled tapes and input to temporary computer storage by special programs written in terms of the addressing system.

Sincerely,

John

John A. Creager

JAC/mak

cc: D. Tiedeman
J. Kaufman

AMERICAN COUNCIL ON EDUCATION
ONE DUPONT CIRCLE
WASHINGTON, D. C. 20036

OFFICE OF RESEARCH

February 28, 1972

Dr. William G. Conroy, Jr.
Division of Occupational Education
1017 Main Street
Winchester, Massachusetts 01890

Dear Bill:

This is to provide some substantive comment on OP#1 and the sampling task differentiation section of OP#2 in the light of our conference discussions and subsequent rereading. The design of the census level of information indicated in Figure 1 appears satisfactory for accomplishing its 3-fold purpose stated on page 2.

In regard to the sample information system, let me first explicitly distinguish (as you already have) the sampling design and logistics from the types of data to be collected on the sample. I would start sampling design by taking all the schools in the "universe" and forming subuniverses by "school types", treating each as a separate subsystem in MISOE development and for sampling purposes. The various types contain different numbers of schools and therefore provide different degrees of flexibility in developing samples. Regarding the school types, the secondary school constitute the largest and most clearly defined group and will be used to comment on further sampling issues. The proprietary schools are probably so different from the public schools, e.g., in the ease with which you will be able to obtain cooperation and possibly in some more substantive matters, that you may want to treat this as a separate school type; a small number of such schools will either preclude doing this or will limit the degree to which finer subsampling of programs and students may be achieved. Perhaps, to a lesser degree, community colleges and "schools" with adult or MDTA occupational educational programs as "school types" will be subject to similar considerations. Here it is desirable to have counts for the whole state to aid in judging feasibility in delineating more detailed sampling plans.

In the secondary school sector with some 1800 schools as the universe base, I would sort these into, say, four geopolitical groups, e.g., metropolitan Boston area, eastern "rural", western cities and towns, and western "rural". From your knowledge of the population density distribution and of the geographic distribution of secondary schools some reasonable definition of these categories should be possible. There is nothing sacred about either the number or labels on these categories; however, an increase in the number will provide and create subsequent problems.

Within the geopolitical categories, a decision is needed as to whether to sample LEA's or individual schools; at least in the larger communities, the LEA may be the central agency covering two or more secondary schools. The advantages of sampling LEA's and including all schools under a sampled LEA are:

1. direct meeting of the requirement that any LEA can be identified with its geopolitical category

2. logistic convenience and possibly lower costs of data collection

3. a built in tendency to sample students in accordance with population density patterns.

The disadvantage is that each school in the state does not have an equal opportunity to be represented in the sample, but proper weighting of data in estimating population totals can allow for disproportionate, random sampling within cells of the sampling design.

Insofar as the LEA's may cover more than one school type, you may want to take logistic advantage of that fact and coordinate the sampling of the other school types with that of the secondary schools. This notion implies nonrandomness in the sampling of the school types unless they exist in sufficient numbers such that the set of schools in the sampling cells can be subdivided into those so coordinated with the secondary school sampling and those which are not. This would still not please a pure mathematical statistician but may be worth considering if the "counts" are favorable and logistic convenience is a strong trade-off point.

Taking all secondary schools within sampled LEA's (except taking a maximum of three if any have more than three), one could take all programs as the next sampling level and all students within programs up to a maximum by grade level. To ascertain the feasibility of this approach and to determine what modifications are required to meet cost and logistic constraints, all readily retrievable data on the enrollments in all kinds of programs in secondary schools should be examined, as well as their "geopolitical" distribution. One would also need to check the U-B-O picture for each program to ensure that the kind of summaries in terms of inter-school similarities shown in your product data example will be possible. One would also want to ensure that the rare programs or those with unique objectives were represented in the sample. Your census data plan should provide the data necessary for sample planning, but preliminary counts, even guesstimates may have to be used, if you must draw samples before the first census implementation and with the idea of later adjustments to the sampling. If the latter contingency can be avoided, fine.

In conference I raised the question of multiple samples -- possibly overlapping -- so that no LEA caries a full and continuing burden of data collection and reporting, and so that a lost LEA (from some logistic goof or refusal to cooperate) can be readily replaced. To this I add two thoughts: resampling every nth year to take advantage of system changes shown by your census data and the possibility of a modified (simplified?) hierarchy of stratification for expenditure data. To ensure a tight linkage between expenditure and other data, information should be obtained from the same ultimate sampling units, so I am having second thoughts about separate samples for detailed expenditure data. Also as indicated in conference, comparisons of LEA information on inputs, processes, and products, on the sample may be summarized within stratification cells, without weighting the data, but comparisons with data summarized across cells, or aggregated at the state level for estimates of census parameters in various subsystems will require data weighting under such procedures as outlined above.

In OP#2 your delineation of the sample task seems generally consistent with the above, but other issues are raised to which I should respond. First, the sample-population relationships provide no serious problem if sampling within the hierarchical cell structure is random and if weights can be computed from the census data, as appears to be the case. The adequacy of weighting any data not obtained in the census will depend on the correlation between those sample data items and the census data items in which weights are based (and, of course, with the stratification variables). There is a subtask on weighting and parameter estimation which will require further discussion; for now, I believe a reasonable weighting and estimation scheme is possible with the proposed census-sample delineations discussed to date.

In discussing "camera effects", it is my judgement that you may be giving too much weight to this possibility at the expense of your other consideration of implications for analysis. For sampling purposes, each year's census cohort should be considered as the population giving rise to a cohort sample. For time trends analysis, given adequate annual cohort sampling, real temporal changes in population parameters should be estimable from weighted samples. If you select new samples between the end of process and product and between product and impact, there will be no basis for matching data on individuals for correlational analysis across these very important longitudinal points. You will, of course, be able to make descriptive summaries within program types for each data element and some limited kinds of cross-sectional things. In programs like Project TALENT and ACE program, we emphasize the followup of a cohort sample and retain matchability in the longitudinal data. This area needs more discussion because both David and Jack have expertise in these matters. It may be that the testing effects are more serious with the psychomotor and other variables you are considering than those we have been working with. You must also consider whether you have enough units within the sampling design to proliferate additional samples ad infinitum.

I turn now to some issues raised on OP#1 regarding data types. The basic IIPI element structure is "right on". In my next communication, I plan to comment on space differentiations more fully. In the process space, I would certainly encourage the idea of picking up sequencing information wherever you can find variations within programs across schools, even going so far as to ensure that this is represented in the sample. The hypothesis that this is an important factor in attainment of objectives is an important one.

Your question of identifying process elements of a program accounting for product data variance is one calling for the regression model; ditto for impact data variance and that is why I am concerned about matching capability in the longitudinal data. In order to deal with prediction of configurations of objectives some decision will be required about how they are to be weighted and combined in analysis: e.g., one could unit weight then, thus treating all as equally desirable, equally important, and having costs and benefits (don't you believe it!) -- or one can assign weights to objectives that allow variations in these matters. I would avoid the classical canonical regression model which would assign weights to maximize predictability of the resulting configuration and such weights are not necessarily the most relevant. For this reason, I see no hurry about having a canonical regression capability in your computer software unless discriminant analyses are anticipated. The above line of reasoning assumes, that once weights are assigned to objectives, the criterion collapses into a single composite variable. This may not be the best way to operate on predicting configurations, but it occurs to me that once MISOE is operational, some of the cost-benefit and impact information can be fedback to provide improved weighting schemes for defining configurations. Another thought is to group configurations and use discriminant functions to predict which subjects are most likely to belong to which class of configurations. Considering that a program with 10 objectives, achieved or not, would have 2^{10} configurations, this approach seems at first sight to be a formidable one. But in some ways, I find a greater intuitive appeal to discriminating configural attainment than predicting some weighted average of configurations.

If an LEA experiments with a process change in a certain program, a comparison of multiple regressions of product variables on old and new sets of process variables, combined with data on the changes in percentages of students attaining objectives will give a provisional answer. However, the data may be found in a school not in the formal sample, being picked up in the census data, and because it is found in a single school, be available on a small sample. It would need to be checked (cross-validated) on another group going through the program and the census data examined to see if other schools are trying the same changes.

I wish to repeat and expand a little on a remark I made in conference about input data. In order to control predictions of product and impact data from process variables for differential input, you will need to include some measures of prior exposure to experiences affecting performance on the objectives. It is not practical, perhaps impossible to pretest all sample students on all objectives, but some, even crude elicitation of prior work after school or on weekends in a garage (for auto mechanics) or in a beauty shop (for cosmetologists) should be devised and obtained. It is conceivable that such experiences may be going on concurrently with the program process, thus contaminating the process space effects. This may not be undesirable for achievement of the objectives, but should not be treated as a process effect unless it is aided, abetted, and otherwise officially part of the process. Logistically, information on this needs to be obtained during or at the end of process, remembering to treat the data in analysis as control variables, not properly part of either process or product space.

The issue you raise (OP#1, page 10) about differential treatments for different groups of students is one under continuing discussion in the methodological literature, involving heterogeneity of regression, moderator variables, etc. Either Dave or I can give you many references. However, what appears to be a recent breakthrough has been developed by Don Rock and his associates at Educational Testing Service in Princeton, New Jersey, and is reported in the current issue of the American Educational Research Journal. It involves a strategy requiring computer software in regression, hierarchical grouping, and discriminant analysis. Regarding the outcome probability tables you mentioned, I enclose my "dream" paper. Some agencies (e.g., American College Testing) are actually doing this kind of thing, perhaps prematurely.

I defer to Jack's expertise on ways of obtaining cost data. I should not think it necessary to get detailed cost data on the census basis, but enough gross data to permit weighting more detailed sample data. I suspect, too, that working with clusters of objectives and meeting other difficulties discussed may require some application of hierarchical grouping of programs and objectives.

Both Dave and I have had considerable experience with the logistic problem of obtaining followup data after program completion. I agree with your idea of getting general information on an actuarial basis and supplementing this with greater depth "clinical" information on a small group. I would, however, give lower priority to the latter. In regard to the sampling plan for followups, I suggest:

1. followup all sample subjects coming out of small-enrollment programs, but cut costs by taking random samples of those coming out of larger programs. This implies the need for adjusted weights on the followup data; no serious problem.

2. be prepared to perform at least one wave of followup of nonrespondents. The fact that you will have extensive input data on the subjects will provide information on the nonrespondents and a basis for computing adjusting weights for nonresponse bias in longitudinal data. Our experience has been that females are more likely than males to respond to mail followups, whites than blacks, and "brights" than "dulls". I understand from John Flanagan that he is interested in these matters and doing empirical studies bearing on them. Dave will be able to give you more on this.

3. it is crucial, if you plan to followup by mail, that viable addresses be obtained and maintained. We have found the student's home address very useful since his family often forward mail to him. This information should be obtained on input for the full sample and maintained as a confidential name and address file in your shop. I can provide some literature from our shop on mailout bias control techniques and on confidentiality issues.

I have little comment on the analytical data types beyond an appreciation for the plan to code product data to both USOE and DOT codes and to note that most of the issues raised should be tractable when the IPPI space differentiations and instrumentations can be implemented. I look for enlightenment on ways of establishing dollar equivalence of non-economic outcome variables. Jack has mentioned some general principles which if properly applied seem critical to the success of MISOE. Some "control group" data are available in both Project TALENT and in the ACE program. In neither case are they as tightly linked to cost-benefit data as we would like.

I promised Elizabeth Weinberger some army and navy contacts, parallel to the air force contact I gave her about occupational analysis, thinking that this would be helpful sources of information in her instrumentation problems, and so it might still be. I learned that there is much larger effort going on in DOD with interservice coordinations which involves computerized occupational data systems for management, training and manpower. I think you may wish to explore what they are doing and decide what aspects are most useful to you and your staff. I gather that the coordination is from the office of General Platt, Assistant Secretary for Manpower and Reserve Affairs, Director of Utilization in the Pentagon, and that the best contact at the level is a civilian, Mr. Robert Groover, in charge of the Occupational Information Service Center: Phone (202)-697-8244. The Dr. Raymond Christal whose name and address I gave to Elizabeth and his colleague Dr. Robert A. Bottenberg at Lackland AFB developed a computer occupational data analysis program (CODAP) which is used in the Air Force and I believe, by the Marines. Christal and Bottenberg each gave papers at a NATO conference held last year in Cambridge and I believe getting copies of their papers may be useful to you.

The army counterpart of Dr. Christal turns out to be Dr. Cecil Johnson in the Behavior and Systems Research Laboratory, located in the Commonwealth Building at 1300 Wilson Blvd. in Rosslyn, Virginia. I understand that the 202 area code can be used for any DOD branches in the area, even those located in Virginia. If you have any difficulty, call DOD central operator, 202-545-6700.

Mr. Groover just returned my call and gave me the names and phone numbers of the other service branches key persons:

Navy: Commander Bruce Cormack of the Canadian Armed Forces on a tour of duty with the USN, has two offices. One at Bolling AFB, phone: OX-3-2712. The other is in the Personnel Research Division of BuNavers which is in the process of moving this weekend to the Arlington naval annex on Columbia Pike, the new number being OX-4-5626. I understand that a Dr. Ballard in that unit is also knowledgeable about the naval activity in this area.

USMC: Col. George Caradakis, Company D, Hdq. Battalion, Marine Corps Base, Quantico, Virginia (703-640-2890?).

U.S. Coast Guard: Mr. Joe Cowan, (202)-426-0891, in the Psychological Research Branch (P-1), U.S. Coast Guard Headquarters, 400 Seventh Street, S.W., Washington, D.C. 20590.

I also promised to send Martin Breslow some information about our statistical computer package. On discussion with our data processing chief, I learned that some difficult legal and other hassles would develop if we were to try to give you the package itself, and we are out of our (outdated) manual. In lieu of this, you should contact David Armour at the Harvard Computer Center for information about the Fortran version of Data Text which he is developing and, is about ready for use. A Data Text Primer is available from him I understand, for $5.00, and probably is the best thing to start with, before deciding whether to negotiate for a copy of the package or to develop a modest version in-house. Our system is an adaptation of an older version of Data Text and was rather costly to adapt and convert.

Sincerely,

John

John A. Creager
Research Associate

Enclosure
cc: Jacob J. Kaufman
David V. Tiedeman